Lifestreams: a modular sense-making toolset for identifying important patterns from everyday life

Cheng-Kang Hsieh[1], Hongsuda Tangmunarunkit[1], Faisal Alquaddoomi[1], John Jenkins[2], Jinha Kang[1], Cameron Ketcham[2], Brent Longstaff[1], Joshua Selsky[2], Betta Dawson[1], Dallas Swendeman[3,4], Deborah Estrin[1,2], Nithya Ramanathan[1,5]

UCLA Computer Science Department[1], Cornell Tech[2], UCLA Global Center for Children and Families[3], UCLA David Geffen School of Medicine Department of Psychiatry and Biobehavioral Sciences[4], Nexleaf Analytics[5]

changun@cs.ucla.edu, hongsudt@ucla.edu, faisal@cs.ucla.edu, jojenki@cornell.edu, jinhakang@ucla.edu, cketcham@cornell.edu, blongstaff@ucla.edu, joshua.selsky@cornell.edu, bettad@cens.ucla.edu, dswendeman@mednet.ucla.edu, destrin@cs.cornell.edu, nithya@nexleaf.org.

ABSTRACT

Smartphones can capture diverse spatio-temporal data about an individual; including both intermittent self-report, and continuous passive data collection from onboard sensors and applications. The resulting **personal data streams** can support powerful inference about the user's state, behavior, well-being and environment. However making sense and acting on these multi-dimensional, heterogeneous data streams requires iterative and intensive exploration of the datasets, and development of customized analysis techniques that are appropriate for a particular health domain.

Lifestreams is a modular and extensible open-source data analysis stack designed to facilitate the exploration and evaluation of personal data stream sense-making. Lifestreams analysis modules include: feature extraction from raw data; feature selection; pattern and trend inference; and interactive visualization. The system was iteratively designed during a 6-month pilot in which 44 young mothers used an open-source participatory mHealth platform to record both self-report and passive data about their diet, stress and exercise. Feedback as participants and the study coordinator attempted to use the Lifestreams dashboard to make sense of their data collected during this intensive study were critical inputs into the design process. In order to explore the generality and extensibility of Lifestreams pipeline, it was then applied to two additional studies with different datasets, including a continuous stream of audio data, self-report data, and mobile system analytics. In all three studies, Lifestreams' integrated analysis pipeline was able to identify key behaviors and trends in the data that were not otherwise identified by participants.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences – *Psychology and Sociology*.

SenSys '13, November 11 - 15 2013, Roma, Italy

General Terms

Design, Experimentation.

Keywords

Mobile Health, Mobile Systems, Personal Data Analysis.

1. INTRODUCTION

The use of personal data streams to improve health (mHealth [10, 16, 23, 34]) incorporates a variety of methods utilizing smartphones, including passive collection of activity, location, and communication; prompted self-report (e.g., experience sampling [21]); and usage data from health *apps* (e.g., PTSD Coach, WellDoc [45, 46]). Such detailed data collection can be used to systematically monitor chronic conditions and health behaviors outside the clinical setting for both research and intervention.

In the raw, these data are highly variable and difficult to interpret. However, researchers across many health domains are working to develop appropriate and customized analytics, to transform these multi-dimensional, heterogeneous data streams into actionable and robust **behavioral-indicators**. The hope is that such behavioral indicators can be used to characterize a user's baseline, and then to identify significant variations, trends and shifts in specific behaviors or symptoms that are relevant to an individual's behavior and health.

Lifestreams is an analytical software stack that facilitates the iterative exploration process needed to define and evaluate specific behavioral indicators. The Lifestreams stack consists of four layers—*feature extraction and aggregation, feature selection, inference and visualization*. Each layer in Lifestreams consists of modular building blocks that can process the data provided by the lower layer, and then send the result to the next layer up.

In this paper we demonstrate the power of Lifestreams using original data acquired during a 6-month, 44-person, NIH-funded mHealth pilot in which young mothers ran a smartphone application suite that captured passive mobility data and self-report about their diet, stress and exercise. Using Lifestreams' data analysis pipeline, we present three sets of analytical results based on inference methods developed to explore participant's behaviors. These inference methods were developed in an iterative fashion (in collaboration with subject matter experts) using the Lifestreams stack to identify significant behavioral patterns in specific participants. The accuracy and utility of the analytical building

blocks were qualitatively verified through interviews with the participants; Lifestreams visualization was used to guide some of these discussions.

In summary, the key contributions of this paper are:

- Lifestreams, a 4-tiered software stack and extensible set of analytical building blocks that facilitates the exploration of diverse personal data streams to extract important behavioral indicators (Section 2).
- Demonstration of Lifestreams' ability to analyze multiple heterogeneous data streams to highlight significant correlations and changes over time even across multiple simultaneous behaviors during a real-world 6-month study with 44 young mothers; and includes a qualitative study based on interviews with 8 out of 44 moms who were able to make use of the Lifestreams visualization component guided by a research coordinator trained in the system (Section 3).
- In order to explore the generality and extensibility of the Lifestreams, we applied it in two additional contexts, one with continuous audio and another of larger scale (Section 4).

2. Lifestreams: A Modular mHealth Data Analysis Stack

Lifestreams is an open-source data analysis stack designed to run on top of an existing mobile data collection system [1]. The Lifestreams stack consists of four layers—*feature extraction and aggregation, feature selection, inference and visualization*. Each layer consists of modular plugins or building blocks that can process the data provided by the lower layer, and then send the result to the next layer up in the stack. Figure 1 shows a diagram of the Lifestreams analysis software stack.

- At the bottom of the stack are different personal data streams. For our deployment and analysis, we used data collected by *ohmage*, an open-source personal data collection platform (Appendix A). These streams include intermittent self-report data (both prompted and user-initiated) and passive data streams collected from sensors and applications on-board the mobile device (e.g. accelerometer data, location traces, communication app usage, etc.).
- These data streams are then sent to the feature extraction layer, which is the first step in transforming them from raw datasets into actionable behavioral indicators. For example, the raw accelerometer and location streams are passed to an *Activity features* module to be transformed into meaningful classifications of activities (i.e. still, walk, run, drive). After features are extracted, they may be further aggregated based on spatial and temporal attributes.
- Features produced by the feature extraction layer are multi-dimensional with many of them irrelevant or redundant to the analysis goal, and therefore need to be further processed by the *feature selection* layer for dimensionality reduction.
- The *inference layer* uses the selected features to detect patterns and trends, and infer behavioral states.
- *Visualizations*, tightly coupled with the analytical building blocks, make the analyses available and actionable for the end user.

[1] Lifestreams is open-sourced under Apache License Ver. 2.0. The source code is hosted on https://github.com/changun/Lifestreams/.

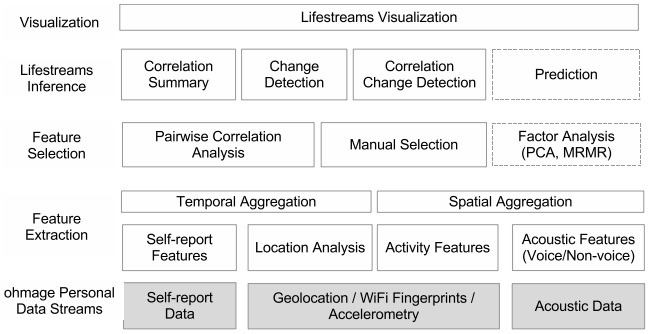


Figure 1. Lifestreams Data Analysis Stack

Lifestreams is designed for behavioral and health sciences researchers (our end users) to analyze high-dimensional mHealth datasets collected from study participants. Their research objectives are to gain insight into a particular condition, or to design and evaluate the efficacy of an intervention. In this paper, most of our analysis focuses on diet, stress, and exercise data collected during an NIH-funded study with 56 Moms, 44 of whom carried mobile phones running ohmage. Data streams were captured via self-report and passive accelerometry and location traces during the 6 months of the study by the Moms; the study is described further in Section 3. Lifestreams' existing analysis modules are not meant to be comprehensive. They provide an initial set of techniques that allow us, and the broader community, to start our data exploration and to validate the techniques and findings. We focus on modularity so that Lifestreams can be extended and generalized to a wide range of research studies with varying types of data streams and research goals.

In the remaining part of this section, we describe each layer of the Lifestreams' data analysis stack.

2.1 Feature extraction & aggregation

We use the term feature extraction to refer to a process of transforming input data into a representative set of features that describe the data and could be used to calculate a behavioral indicator. For example, the activity feature modules described below produces a stream of important features related to a user's activities throughout the day, using the accelerometer, Wi-Fi signature, and location data streams. Through collaborations with clinicians and patients, we have identified several disease domains in which activity and location patterns are relevant to determining the patients' status. These diseases include depression, chronic pain, gastro-intestinal, inflammatory, and auto-immune disorders. For each disease and demographic in which such measures might be used, domain-specific exploration and knowledge is essential. Here we focus on the system support needed to define and refine the use of this type of data for a specific study or intervention. In the following sections, we introduce each Lifestreams' feature extraction module in more detail.

2.1.1 Activity features

Lifestreams currently implements the following activity features:

- Mobility state per minute. States are *still, walk, run,* and *drive;* each minute of a stream is annotated with the associated mobility state produced by the classifier. While it is

Table 1. Five different measures used to compute correlation coefficients between quantitative, ordinal, and nominal features.

	Quantitative	Ordinal	Nominal
Quantitative	Pearson's r	Spearman's rho	Point Biserial rpb
Ordinal	Spearman's rho	Spearman's rho	Rank Bisereal rrb
Nominal	Point Biserial rpb	Rank Bisereal rrb	Phi

susceptible to noise from individual classifier errors, it allows short events on the order of minutes (not seconds) to be identified robustly, e.g., a walk to the water cooler or waiting for a long stop light.

- Mobility events: Mobility events are a higher-level set of longer periods of activities, providing an overview of the general activity level for the day. This gives a better indication of how sedentary and active users were, and when they had to travel. This feature includes start and stop time for the event and the associated mobility state.
- Total Distance Travelled: Each mobility event is also annotated with the distance travelled during that event. Total distance traveled is used to determine how strenuous an ambulatory activity is, or how long a drive is. This can help determine activity levels or how much the user had to travel.
- Geo-Diameter: The longest distance between two locations included in a mobility event. Geo-Diameter is useful for determining how far from home a user traveled during the day, and offers additional information to the health researcher beyond total distance traveled (see [44] for a sample application of geo-diameter).

Activity events in ohmage are calculated using accelerometer, Wi-Fi fingerprints, and location traces collected on the phone by ohmage. The classification is based on supervised machine learning techniques [43]. The activity classification is performed in two stages. First, the accelerometer records one second's worth of triaxial samples. To be orientation-independent, it first calculates the magnitude of acceleration and then calculates the variance and FFT coefficients on the set of values. These are used to classify ambulation, i.e., sedentary, walking, or running. If the user is classified as sedentary, then the current Wi-Fi fingerprint is compared to the access points encountered in the past 10 minutes. If the fingerprint has not changed, the user is classified as still. If it has changed, the user is presumed to be driving. If there are not enough Wi-Fi access points for the comparison, then GPS speed is used to determine whether the user is driving or stationary. This specific activity classifier is modular and can be replaced by other classification techniques, such as in [32] and [35].

Our activity classifier produces a series of activity estimates on the order of one per minute. Since user routines consist of blocks of activities, we calculate activity events, that is, periods of time during which the user was in a single activity. To prevent noise resulting from occasional miscalculations, we smooth the data, ignoring brief changes that revert to the previous activity, such as one instance of walking during a period when the user is still. There is an obvious tradeoff between responsiveness to short events and noise reduction; features can be calculated with various degrees of smoothing depending on the objective of the study and the noisiness of the data source.

Once periods of activity are identified, features of the user's routine are calculated from event properties, such as the start and end times and duration of the event, and the number of events occurring within a time window. Features are also calculated based on

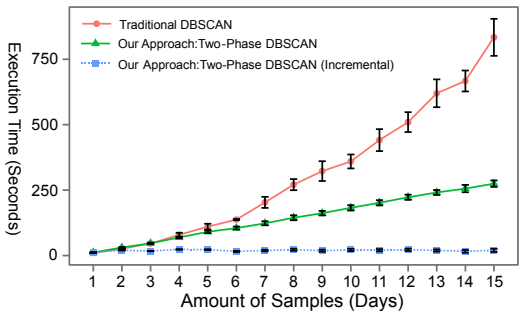


Figure 2 Comparison of execution time between the traditional DBSCAN and our Two-Phase approaches for different amount of samples (sample rate=1/60Hz). Each point is an average of 10 runs. The error bars are the corresponding 95% confidence interval.

alternate activity labels. Rather than handling each activity class separately, we aggregate results of the same type. For example, total cumulative time spent sedentary or in physical activity might be used as a behavioral indicator in some contexts. The features largely depend on the domain and how they are used. For example, if it does not matter whether the user is sedentary in a moving vehicle or in a chair at home, the Lifestreams user can aggregate drive and still into a single activity class, sedentary.

2.1.2 Semantic Location features

Lifestreams also produces a stream of important features related to a user's *semantic* location throughout the study [22]. Each location data point collected by ohmage contains a Wi-Fi fingerprint and a GPS coordinate. From this data stream we extract features, including: (a) Places: Places are potentially meaningful locations to a participant. (b) Duration at a Place: The amount of time that one stays at a place. (c) Time of Arriving and Leaving a Place. (d) Daily Work Schedule Deviation: The deviation of a participant's daily work schedule from her normal working routines. This feature is calculated by the time difference between the earliest work arrival time of that day and the median of earliest work arrival time of all the working days during the study period. Such deviations have been shown by behavioral research to correlate with patient's physical, psychological, and social outcomes [39].

A common approach to discover meaningful places from passive location traces is to assume that *a place is a location where the user stays for more than a certain period of time* [4]; such locations are revealed with clustering algorithms, where each resulting cluster is regarded as a place. A density-based clustering algorithm, called DBSCAN, is commonly used for this purpose [9]. DBSCAN defines two points as "neighbors" if their distance is below a threshold, referred to as *Eps*. If a point has more than *MinPts* of neighbors (i.e. its local density is high), this point and its neighbors are declared as a cluster. Two overlapping clusters will be merged into one larger cluster recursively until no cluster is overlapped. The advantages of DBSCAN are three-fold: (1) it can reveal places of arbitrary shapes; (2) it can work on location traces with arbitrary number of places; and (3) its final clustering result does not depend on the initial random assignment of the clusters and is more reliable [9].

However, DBSCAN's complexity is $O(n^2)$, where n is the number of data points, and therefore does not scale well to location traces collected in real-world studies where n could be quite large. To address this problem, instead of extracting places from all data

points at once, we adopted a two-phase process to first extract daily places based on 24 hours of location traces and then merged similar places extracted throughout the study into a final set of places. DBSCAN is used to identify places in both phases. Phase 1 DBSCAN extracts the ambient signal signatures that characterize the Wi-Fi and GPS signals of each place, and Phase 2 DBSCAN merges the places based on the similarity of their ambient signal signatures and generates a final set of places.

The time complexity of our two-phase algorithm is $O(K^2D+M^2)$; where K denotes the number of data points collected daily, D denotes the number of days, and M denotes the total number of places generated in the first phase. As shown in Figure 2, when the study duration increases from one day to fifteen days, the execution time of our Two-Phase DBSCAN approach increases 24.6x as compared to 92.6x of the traditional DBSCAN. More importantly, our approach can run incrementally (referred to as Two-Phase DBSCAN (incremental) in Figure 2). If all places from previous days are cached, we only have to run the first phase for those days that have not been processed, and re-run the second phase to update the final place assignments. These advantages make our approach more appropriate for processing evolving location traces for mHealth studies.

2.1.3 Features extracted from self-report data

Unlike the feature extraction modules that extract features from passive data streams, the module that extracts self-report data features results in a dimensionality expansion from the original dataset.

Self-report data, captured in the form of surveys, are classified into one of the following data types: (1) quantitative (e.g. number), (2) text, and (3) categorical (e.g. single choice/multiple choices). The quantitative data type is used as is. We have not yet implemented text feature extraction but anticipate future NLP plugins to extract features such as vocabulary difficulty and sentiment analysis. In this section, we describe the transformation of the categorical data type which is commonly used in survey questions since it is easy for participants to respond.

There are two types of categorical features — *ordinal and nominal*. For an ordinal feature (e.g. food quality rating as high, medium, low options), a rank or number is assigned to each element to indicate the order among categorical elements. These assignments are usually done by the study researcher and the ordinal values are derived from the study configuration. For a nominal feature (e.g. cause of stress as relationship, financial, school/work, etc. options), all options are independent and have no ranking among them. In our analysis, each option is transformed into a binary feature with the value equal to one if the option is selected and zero otherwise.

For a set of features that are possibly correlated, Principle Component Analysis (PCA) can reduce them to a smaller set of features. In our analysis, we found high correlation among the ten options of the mood questions including happy, calm, stress, etc. These options are highly correlated as both positive and negative mood are not expected to be present simultaneously. Therefore, PCA is applied to transform these features into a single feature that will have a higher value when more positive mood options are selected, and vice versa.

2.1.4 Data aggregation

All features derived from different modules can be further aggregated, when appropriate, based on statistical functions (e.g., min, max, or average), time (e.g. daily or weekly) and/or location. For example, all extracted features can be aggregated over an arbitrary time period and starting timestamp, both of which are adjustable parameters. For our analysis, we aggregate these features over 24 hour periods starting at midnight on the first day of the study to represent daily data points.

2.2 Feature selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. Redundant or irrelevant features are removed during the feature selection process. Feature extraction creates new features from the raw datasets, while feature selection returns a subset of these features. We describe a simple feature selection module based on pairwise correlation analysis. Features with high correlation, using pairwise correlation analysis, are then highlighted. Other feature selection modules can be substituted, such as principal component analysis (PCA), and minimum Redundancy Maximum Relevance (mRMR) [30].

2.2.1 Pairwise correlation analysis

A correlation matrix is one of the most common methods used to summarize relationships between any two features. However, when dealing with a multi-dimensional heterogeneous dataset where different features are of different types, one correlation measure cannot simply be applied to all features. Five different correlation measures are used to compute the correlation coefficients between quantitative, ordinal, and binary (nominal) features (see Table 1).

The proper hypothesis test of each correlation measure is performed with no correlation as the null hypothesis. A desired significance level (with 0.05 as a default) can be specified. Any correlation that does not have sufficient significance level will be excluded. The pairwise correlation coefficients by default are calculated across all data. However, the user can choose to calculate correlation coefficients on subsets of the data through an interactive interface (described in Section 2.4); for example, based only on weekday or weekend data to investigate weekday/weekend effect.

2.3 Lifestreams Inference

Inference is a process of arriving at some logical conclusions from premises known or assumed to be true [1]. The Lifestreams inference layer focuses on techniques that enable researchers to understand user contexts, behaviors, and changes, based on selected features. We provide three different inference methods to investigate long-term relationships between participants' behaviors and to detect behavior changes from the participants' behavioral data streams.

2.3.1 Correlation summary

The pairwise correlation coefficient is useful for exploring relationships or dependencies between any two sets of time-series data. A strong and significant correlation indicates consistent relationship throughout the entire time-series. The pairwise correlation analysis module calculates a three-dimensional (3D) correlation matrix across all pairwise features and participants. However, the 3D matrix is large and hard to visualize. The correlation summary module is a thin layer on top of the pairwise correlation analysis that provides different capabilities to quickly shift through the large 3D data and draw conclusions about long-term relationships, such as pairwise correlation for an individual and similarity across individuals.

The 3D matrix can be filtered by individual participant and by a threshold value to produce a 2D correlation matrix for that individual. The 2D matrix shows only features with correlation coefficient values stronger than the specified threshold (e.g. 0.3). This capability allows researchers to quickly explore pairwise

features with different coefficient values. A statistical summary matrix capturing similarity across all participants based on different coefficient thresholds can be calculated. This matrix shows, for each pair of features, the number of participants with the correlation coefficient greater than or equal to a specified threshold. We can further filter this matrix by only showing pairwise features with the number of participants greater than or equal to a specified value (e.g. 25%). This capability will allow researchers to identify common pairwise relationships among participants. We show examples of these plots based on our pilot data in Section 3.3.

2.3.2 Change detection (single feature)

The ability to detect changes in an individual's behaviors will enable behavioral and health sciences researchers, as well as context-aware applications, to react accordingly and to provide active-interventions to participants. For example, a mobile app could display stress reduction techniques to a participant when changes that have previously-resulted in higher-level stress are detected. In this section, we introduce the change detection module that detects behavior changes on each individual feature. We will introduce another change detection module that detects the correlation changes between pairwise features in Section 2.3.3.

We model the behavior change detection problem as a statistical change detection problem. We detect behavior changes by comparing the distributions of recent observations of the participant's behavior with previous observations, and identifies a potential behavior change when these two distributions are significantly different. Such statistical change detection approaches have been applied in many fields, such as finance, robotics, and quality control [19, 24].

Our iterative procedure for behavior change detection algorithm works as follows given a sequence of observations $x_1, \ldots, x_t, \ldots$, where x_t is the observation at time t. Suppose we want to know if any behavior changes have occurred before time t. We apply a *two-sample test* to every possible splitting point $t' \in (1, t)$ and compute the test statistic $D_{t't}$ that estimates the *degree of difference* between the distributions of $\{x_1, x_2, \ldots x_{t'}\}$ and $\{x_{t'+1}, x_{t'+2}, \ldots x_t\}$. Let t^* denote the time t' that maximizes $D_{t't}$ among all possible t', and compare D_{t^*t}, which is the maximum $D_{t't}$ for all possible splitting point t', with a pre-defined threshold. If D_{t^*t} exceeds the threshold, we declare that a change has occurred at time t^*. After a change point has been identified, the observations before the time t^*, that is, $\{x_1, \ldots x_{t^*}\}$, will be discarded, and the procedure restarts.

The core issue of such a statistical change detection algorithm is to define a test statistic that best suits the characteristics of the observations. Behavioral data is known to be non-normally distributed and might not satisfy the assumptions of many parametric statistics [31]. Therefore, we adopt the Mann–Whitney (MW) test, which is a non-parametric test, to estimate test statistic $D_{t't}$ for quantitative and ordinal behavioral data streams [38]. The MW test performs a two-sample test based on the ranks of observations instead of the raw values. Although it is less sensitive than *Student-t* test, we found the MW test to be a more appropriate choice for behavioral data. For categorical binary data streams, we adopt Fisher's exact test, which is a common test used to detect distribution changes in a Bernoulli sequence [33].

Two parameters have to be determined for this algorithm. *Startup time* is the number of observations after which detection begins; it is necessary since the statistical test has low power for a small sample size. ARL_0 is the expected time difference between a false positive detection and the real change point; the size of ARL_0 determines the threshold levels at different times t. There is a tradeoff between the responsiveness of the detection algorithm and the expected duration of the detected changes. We will investigate this tradeoff in Section 3.4 in more detail based on participants' data.

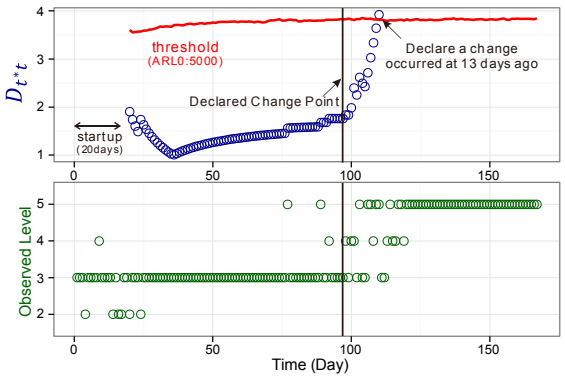


Figure 3. A walkthrough of the change detection algorithm

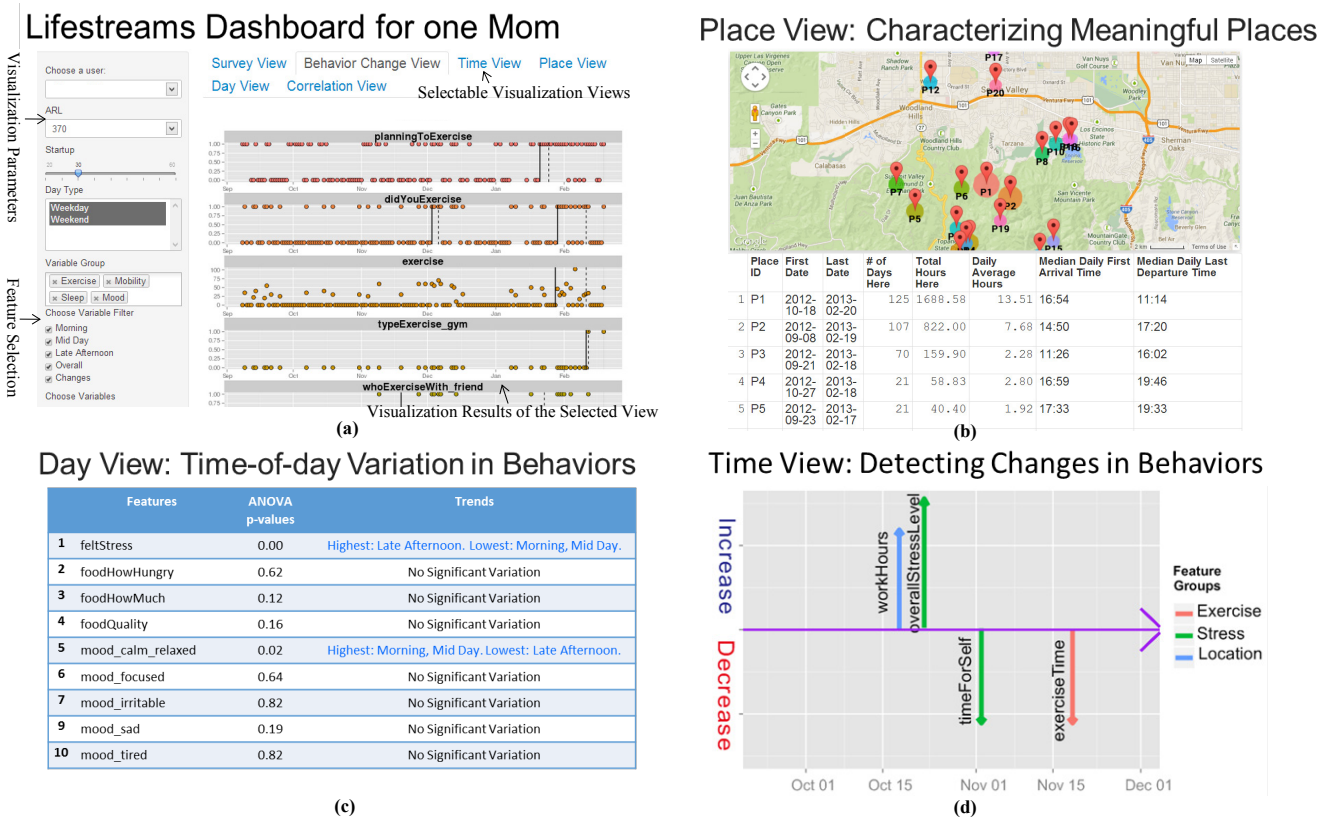
Figure 3 provides a walkthrough of our change detection algorithm. Each green circle on the bottom plot indicates an actual observation on a different *Day t*, and each blue circle on the top plot indicates a computed D_{t^*t} for that day. The algorithm will not start to detect any changes until after *startup time* (20 days). Initially, there are a number of anomalies observed between Day 20 and Day 30. However, these anomalies do not last long and are not consistent enough to trigger change detection. From Day 98, the algorithm starts to observe more consistent higher-level outcomes, and on Day 111, when the computed D_{t^*t} exceeds a specified threshold, the algorithm determines that a change was observed on Day 98.

An R package, called *cpm*, is used to implement the above-mentioned change detection algorithm [33]. Since many individuals have different patterns during the week and weekend, we introduced a weekday/weekend/all parameter (with weekday as a default) to identify different subsets of data for analysis.

2.3.3 Correlation change detection (pairwise)

The pairwise correlation described in Section 2.3.1 is useful for exploring long-term relationships between any two behavioral features. However, the overall pairwise correlation cannot capture the changes of relationships between two features and may overlook important information. The change detection on pairwise correlation addresses this issue. The detection can be used to identify changes in behavior or to discover the shorter-term relationships that could be overlooked in the overall correlation analysis.

We introduce two techniques — pairwise correlation change detection and moving average—to further investigate changes in behavior based on pairwise features. The correlation change detection is modeled in a similar way as the single feature change detection. The main difference is that the test statistic for deriving $D_{t't}$ is calculated based on two sets of observations from the both features. We implemented a detection algorithm developed by Wied et al. [37] to detect the pairwise correlation changes. There are two parameters required for this algorithm: (1) start time m assumes that the correlation will remain constant among the first m observations. (2) r is a parameter that determines the threshold function. A higher value of r results in a higher false alarm rate, but with shorter detection delay. We select the parameter values $m = 20$ and $r = 0$ as a default for a longer detection time, but lower false positive rate.



Place ID	First Date	Last Date	# of Days Here	Total Hours Here	Daily Average Hours	Median Daily First Arrival Time	Median Daily Last Departure Time
1 P1	2012-10-18	2013-02-20	125	1688.58	13.51	16:54	11:14
2 P2	2012-09-08	2013-02-19	107	822.00	7.68	14:50	17:20
3 P3	2012-09-21	2013-02-18	70	159.90	2.28	11:26	16:02
4 P4	2012-10-27	2013-02-18	21	58.83	2.80	16:59	19:46
5 P5	2012-09-23	2013-02-17	21	40.40	1.92	17:33	19:33

	Features	ANOVA p-values	Trends
1	feltStress	0.00	Highest: Late Afternoon. Lowest: Morning, Mid Day.
2	foodHowHungry	0.62	No Significant Variation
3	foodHowMuch	0.12	No Significant Variation
4	foodQuality	0.16	No Significant Variation
5	mood_calm_relaxed	0.02	Highest: Morning, Mid Day. Lowest: Late Afternoon.
6	mood_focused	0.64	No Significant Variation
7	mood_irritable	0.82	No Significant Variation
9	mood_sad	0.19	No Significant Variation
10	mood_tired	0.82	No Significant Variation

Figure 4. Lifestreams visualization allows users to interactively select different feature groups, switch between different visualization views, adjust parameters, and export the visualization results. Users can use Lifestreams visualization to quickly and flexibly navigate through vast and high-dimensional personal data streams and create visual aids to guide the discussion with patients or participants. (Note that the geo-information in the *Place View* has been obfuscated to protect the participant's privacy.)

Moving average is a technique commonly used in time-series data analysis to provide an estimate of the trend of the observations [8]. Moving average smooths out short-term fluctuations and reveals the longer-term trends of the data. Each data point is an average of a subset of data extracted around the original raw data point e.g. a subset of 7-day data starting from the current day.

In our analysis, we transform the original dataset using the seven-day moving average before applying the pairwise correlation change detection algorithm. Similar to single feature change detection, a parameter weekday/weekend/all can be specified.

2.4 Lifestreams Visualization

The purpose of Lifestreams visualization is two-fold: (1) to allow researchers to quickly and flexibly navigate through high-dimensional personal data streams by focusing on important trends/patterns, and (2) to explore visual aids for an intervention, such as to guide discussions with patients in clinical or coaching sessions. It allows users to interactively select different feature groups (e.g. diet, stress), switch between different visualization views, adjust visualization parameters, and export the visualization results. We describe a few examples of visualization views below.

(1) *Behavior Change View* displays time-series plots of individual features and allows users to select only features that contain potential change points during the study (Figure 4a). This function helps researchers select features with potentially-noteworthy trends from a large feature pool.

(2) *Place View* marks the locations that are potentially meaningful to a participant on an interactive map interface (see Figure 4b). These locations are identified by the Location Analysis module described in Section 2.1.2. In interviews with the participants, we found this interactive map, along with the time-related statistics, such as the amount of time spent in each place, and the arrival and departure time, are particularly useful in helping participants to recall the semantic meanings of different places (e.g. home, work place, gym, children's schools, and etc.). This semantic information is in turn used in transforming the participants' raw location traces to more meaningful behavioral features.

(3) *Day View* uses ANalysis Of VAriance techniques (ANOVA) to evaluate how a patient's behaviors vary throughout the day [18]. For example, Figure 4c shows that this participant's stress level and mood tend to have statistically significant variation at different times of day (i.e. the p-value ≤ 0.05). Such analysis is particularly useful when self-report data exhibit high variation and recall biases depending on the hour, such as how much food, how hungry, current stress levels, etc.

(4) *Time View* plots the change points of different features on a single timeline (see Figure 4d). It provides an overview of a participant's behavioral trends and suggests potential interactions among different behaviors (e.g. an increase in stress level shortly after an increase in working hours).

3. EXPLORING A REAL DATASET WITH LIFESTREAMS

In this section, we explore the accuracy and utility of Lifestreams inference methods by applying them to the data collected from 44 Moms who carried smartphones in a 6-month, NIH-funded study of cardiovascular disease risk factors in young mothers. We compare the output from the analyses with qualitative information obtained directly from the participating Moms about their diet, stress, and exercise. We find that in many instances, Lifestreams inference methods were able to automatically identify trends and changes that were confirmed by participants.

These analyses are intended as illustrative examples that could be adapted to the other use cases, such as: to provide feedback for patients in a behavioral intervention program targeting the management of diet, stress, and exercise; to design a visual aid for patient portals or clinical visits; or to provide information to health researchers and coaches about a participant's context and trends in behavior.

3.1 The dataset: diet, stress, and exercise in young moms

In our study, behavioral researchers used Lifestreams to study cardiovascular risk factors in young mothers. The purpose of the study was two-fold. First, to evaluate the validity and reliability of the phone in capturing diet, stress, and exercise for women; and a secondary aim to evaluate the efficacy of using the phone for behavior change. Basic measurement parameters included a participant's daily exercise routines, their diet, and their stress and mood levels throughout the day measured by survey; and a participant's accelerometry, location traces, and mobility states recorded continuously using our automated mobility classifier.

56 young moms residing in the Los Angeles area participated in the study January 2012 - March 2013. Of the 54 moms who completed the study, the experimental group consisted of 44 moms who used our smartphone application to collect data. Of the 44 Moms' ethnicity and races were diverse and their ages range from 18 to 40 with an average of 30 years old. 15 moms worked full time, 24 moms worked part time including studying, and 17 moms worked in the home. All moms had at least one child living at home. The study start date of each mom varied from January in 2012 to September in 2012. 52 moms had completed their study with an average duration of 7 months. 4 moms dropped out in the middle; two moms became pregnant and the other two did not specify. Lifestreams visualization was ready for use in January 2013 for a qualitative study involving 8 out of 44 moms during their in-person interview discussions.

15,599 survey responses were collected across 44 moms in the experimental group. The responses are distributed uniformly across the four surveys, 4,248 (27%) morning surveys, 3,968 (25%) midday surveys, 3,722 (24%) late afternoon surveys, and 3,661 (24%) bedtime surveys. The average of survey numbers answered per user is 354, and participants answered 2 surveys on average per day during periods where participants answered at least once. A total of 115,228 questions were answered, with an average of 2,619 questions per participant and 16 questions per participant per day. The most popular questions participants persistently answered were 'Have you eaten since you completed your last diet survey?' at midday, 'Have you felt stressed in the last two hours?' in the morning, and 'How many hours in total did you sleep last night?'

The participants could choose to turn on or turn off the mobility data collection. 3,834 days of mobility data were collected from 44 moms with a minimum of 5, a maximum of 202, and an average of 87 days. Since the number of the entire study days across all moms is 7,272 days, participants contributed mobility data during one-half of their study period on average.

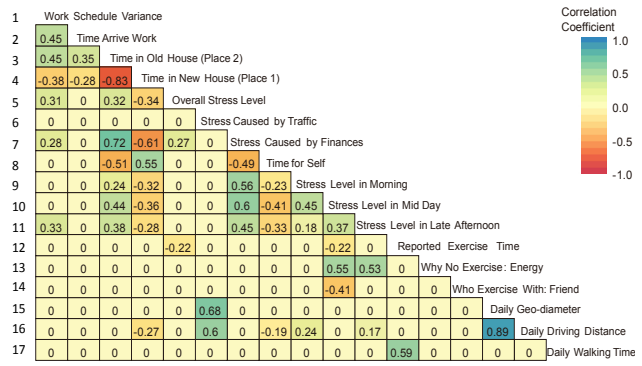


Figure 5. Individual 2D correlation matrix of a participant. This figure shows correlations between the participant's working schedule, stress levels, exercise routines, and daily activity patterns.

3.2 Data analysis results and discussion

The analysis presented here uses self-report survey data and passive accelerometry and location data streams from the Moms. The survey data consists of 47 questions grouped into 4 different surveys for participants to complete in the morning, midday, late afternoon and evening. The self-report data are transformed into extracted features using techniques described in section 2.1.3 including nominal data transformation and PCA. A set of activity and location features was extracted from the passive activity and location data streams. All data are aggregated into a daily summary. At the end of the feature extraction and aggregation process, there are 142 features extracted for each participant.

In the remainder of this section we discuss results derived using several inference modules used in this study. In each section we describe findings for a single exemplar participant in order to highlight the utility of the modules. Agreement between events and relationships identified by the modules and qualitative data from participants indicates the utility of the modules in automatically identifying important features that could not otherwise have been found through manual parsing.

3.3 Correlation Summary

Correlation analysis allows researchers to infer high-level, long-term, behavioral patterns. We use Cohen's conventions to interpret correlation coefficient of behavioral data [31]. A correlation coefficient of .10 is thought to represent a weak or small association; correlation coefficients of 0.30 and 0.50 are considered as moderate and strong correlation, respectively. Only significance levels greater than 0.05 (i.e. p-value $\leq$ 0.05) are considered as statistically significant.

3.3.1 Individual 2D correlation matrix

A participant's 2D correlation matrix among pairwise features can be used to infer her long-term behavioral patterns. Figure 5 shows a specific participant's 2D correlation matrix where only the pairs of features that exhibit correlation greater than or equal to 0.30 are shown.

This matrix shows many interesting relationships between different aspects of the participant's behavior. For example, her daily work

Figure 6. A 3D correlation matrix of all 44 participants. This figure shows an overview of significant patterns among the participants. The number on the left (and right) of each cell indicates the number of participants with negative (and positive) correlation coefficients higher than 0.3.

schedule deviation (1st row) shows moderate correlations with both her daily overall stress level (5th row) and the late afternoon stress level (11th row). Note that the work schedule deviation is defined in Section 2.1.2 and the daily overall stress level is the stress level that a participant rated before bedtime. These correlations between work schedule and the mentioned stress levels were confirmed by the participant—she mentioned that her occasionally-varying work schedule was one of her causes of stress since she could not maintain her regular routines on those days when her work schedule changed. This result also aligns with behavioral studies of working schedule control [39].

The matrix also reveals a strong negative correlation between the daily amount of time she spent in *Place 1* (*new house;* 4th) and *Place 2* (*old house;* 3rd row); which is consistent with the participant moving into a new house during the study period. Moreover, the correlation matrix highlights moderate to strong correlation between her new house and lower stress levels, less concern about finances, and increased time for herself; all of which were confirmed by the participant. In addition, when she had a larger daily geo-diameter (15th row) or a longer driving distance (16th row), her cause of stress was more likely due to the traffic (6th row). Finally, there is a strong correlation between her reported exercise time and the daily walking time detected by our activity classifier. This was also verified by the participant, who reported that her primary exercise activity was walking around the neighborhood.

3.3.2 Correlation statistics across all participants

In addition to the 2D correlation matrix for an individual, a population-wise 3D correlation matrix can provide researchers with an overview of significant patterns among the population, and can also help an individual understand how they fit in with the broader population. Figure 6 shows a selective set of pairwise features that have at least 25% of participants with correlation coefficients higher than 0.3. Note that the number of participants and/or coefficient levels can be changed to different thresholds. For some pairwise features the data show consistent correlation across multiple participants. For example, there are 11 participants that have a negative correlation between whether they exercised or not (2nd row) and daily overall stress level (12th row), and between the time they had to themselves (8th row) and daily overall stress level. Moreover, interestingly, comparing the stress levels at different times of day and the overall stress level, only 8 participants show

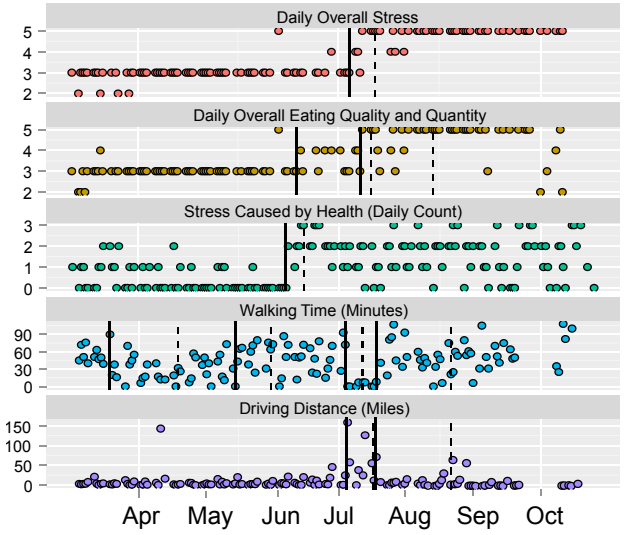


Figure 7. Sample single feature change detection results for a participant (Startup=30 days, ARL_0=500). This figure shows changes detected across daily stress levels, diet, exercise routines, and driving distances.

positive correlation between the reported morning stress level (11th row) and the overall stress level (12th row), while 19 and 20 participants show positive correlation between the reported mid-day stress level (10th row) and the overall stress level, and the late-afternoon stress level (9th row) and the overall stress level, respectively.

In addition, for some pairwise features (e.g. whether a participant exercised vs. planning to exercise), the data show both positive and negative correlation among different participants. For example, for one participant, having a plan to exercise (3rd row) has a negative correlation with whether the actual exercise took place (2nd row).

3.4 Single feature change detection

We found the single feature change detection algorithm (described in Section 2.3.2) to be one of the most useful building blocks in helping researchers make sense of behavioral data. Feedback from participants and the study coordinator was consistent with this finding.

Figure 7 shows sample change detection results for participant 016. The solid vertical lines in the plots indicate the change points estimated by our change detection algorithm, and the dashed lines following the solid vertical lines indicate the time when the corresponding change is detected by the algorithm. Based on the information we learnt in a follow-up interview with the participant, we found that the change detection results were consistent with the participant's real life events. This participant is over-weight and reported that she became more concerned about her health around the beginning of June. This can be seen by the increasing number of times she reported her health as the cause of the stress (3 is the maximum number per day), and the increasing daily overall stress levels, both of which are detected by our algorithm. In addition, she also reported that she stopped eating at restaurants and enrolled in a weight-watcher program, which corresponds to the improvement in food quality detected in June and July. Moreover, the participant reported taking a two-week family trip at the beginning of July, which can be seen from the increase in her driving distance as well as the decrease in the walking time around the same period.

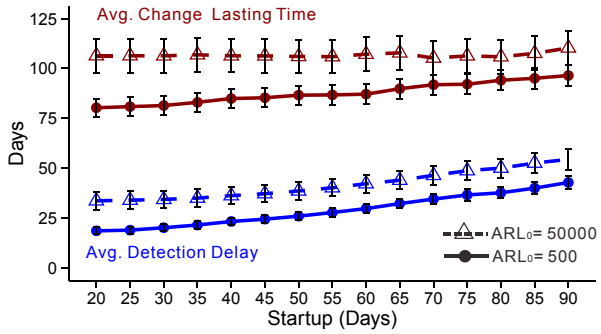


Figure 8. Tradeoff between the responsiveness of change detection and the time the detected changes are expected to last. With a shorter *startup time* or a smaller ARL_0, the algorithm detects changes earlier, but also tends to include temporary changes.

Even though Lifestreams cannot know the specific events that caused the detected changes, the visualization and analytics may help a study coach narrow in on important timeframes and behaviors where significant changes occur. Automated analyses for example, could help healthcare providers give feedback when stress levels begin increasing, and identify the triggers of changes to deploy proper interventions.

As shown in Figure 7, the change detection algorithm usually detects changes within a month after the changes occurred, and some of the changes were temporary (e.g. the increase in driving distance) while some of them lasted longer (e.g. the increase in stress level). There is a tradeoff between the responsiveness of change detection and the expected duration of the detected changes; and this tradeoff can be made by adjusting the two parameters of the change detection described in Section 2.3.2. Figure 8 shows the effects of different values of these two parameters. Each point in the figure is an average of all the changes detected in 142 features of 44 participants, and the error bars are the corresponding 95% confidence intervals. With a shorter startup time or a lower threshold (i.e. lower ARL_0), the algorithm is able to detect changes earlier, but it also results in a shorter average change duration—the average time interval between two consecutive change points. A shorter average change duration indicates that the algorithm detected more short-term changes. The change detection should be adjusted to fit the purposes of different applications. For example, an intervention program might want to detect changes as early as possible to provide participants with timely feedback, but a study on the effect of a new depression treatment might be willing to trade the responsiveness of change detection for a higher probability that the detected change was persistent. The results presented here can be used as a guideline for Lifestreams users to choose the parameters that best fit their purposes.

3.5 Pairwise correlation change detection

Another type of change detection, as described in Section 2.3.3, is to detect changes in the correlation of pairwise features. We applied this technique to all pairwise features of each participant. Figure 9 shows a sample result of correlation change detection for participant 016. As with the change detection, the solid vertical line indicates the estimated change point, and is followed by a dashed line indicating the time when the algorithm can detect the change. In the beginning of the study, her daily stress level was strongly correlated with whether she exercised with her child (with a

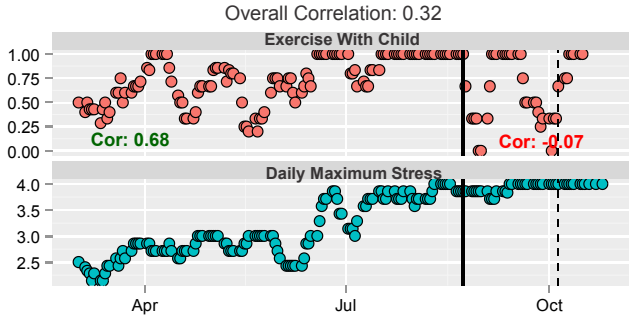


Figure 9. A sample correlation change detection result for a participant. This figure shows how the correlation between two features may vary over time.

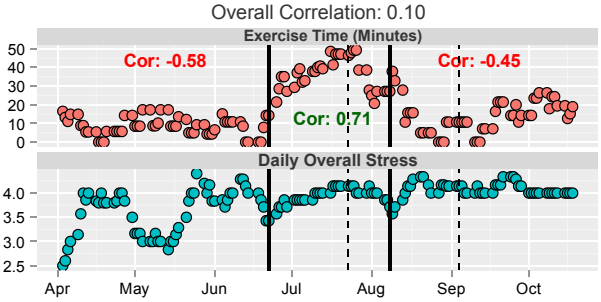


Figure 10. A sample correlation change detection result for the other participant.

correlation of 0.68), but after mid-August, the time when her school starts, the correlation between these two features became 0.07, which is considered as weak or no correlation.

Interestingly, looking at the overall correlation number of the whole study period would not have been useful in detecting the strong relationship between these two features at the beginning of the study. It is only when the time-series is segmented into two components that the pattern during the first three-fourths of the study emerges.

Figure 10 shows another example of correlation changes. It shows that, for participant 029, while the overall correlation between daily exercise time and stress level is weak (0.1), the correlations of different subsets of time are much stronger: The correlation changes from -0.58 before the first change point, to 0.71, and then back to -0.45 after the second change point. The first change timeframe is consistent with the beginning of the summer class session 1 of her school when the participant started to spend more time exercising due to weight concern/stress which then tapered off during summer class session 2 when she started to fall back to the minimum routine exercise required by her position. After mid-summer session 2, her high stress was reported to be due to class work required for advancing in her position and she spent less time exercising due to class load, hence the negative correlation.

4. EXTENSIBILITY AND GENERALIZABILITY OF LIFESTREAMS

This section evaluates the extensibility and generalizability of Lifestreams by demonstrating how Lifestreams can be useful for both smaller preliminary explorations (Family Wellness), and larger more mature studies (Mobilize); they also demonstrate how readily new analyses can be added to Lifestreams and applied to

additional data streams such as audio data and phone usage data that were not originally contained in Lifestreams. In this section, we briefly describe the two studies, their unique requirements, how Lifestreams modules were reused or newly developed to fulfill these requirements, and preliminary findings.

4.1 Family Wellness Study

Family Wellness is a pilot study aimed to evaluate the feasibility of using mobile technologies for familial behavior studies. To date, 15 families have been recruited and completed the study. All families included mothers and one child aged 10-14, and about half also had fathers participating. The participants used ohmage to answer four surveys a day for two weeks about their interactions with each other, and individual-level responses on stress and affect. In addition, the participants ran a smartphone-based audio sensing application that collected and uploaded in situ audio features used to classify speech versus non-speech in the environment.

For this study, we reused the existing self-report module and the correlation analysis module, and developed two additional modules including: 1) an audio data extraction module to extract audio features; and 2) an inter-user correlation module to investigate interactions among family members. 82 lines of code were added to implement these two additional modules while 468 lines of code in pre-existing modules were reused for this study. We report our preliminary analysis results below:

(1) Correlation between the audio data and family interactions. Due to the exploratory nature of the audio data collection module, only one participant's complete audio data was captured and analyzed. For this particular participant, there was significant correlation between speech/non-speech audio features and self-reported family interaction. For example, daily speech time detected by the application is highly correlated with the levels of argument reported with the study child over a day (coefficient = 0.85, p-value < 0.001) and with the levels of argument reported specifically in the evening (coefficient = 0.75, p-value < 0.01). While more data are needed, these preliminary results suggest that there is a potential value in applying audio sensing in such a context, and the researchers are planning a larger scale study.

(2) Discordance among family members' subjective feelings. The inter-user correlation module enables researchers to compare the responses from multiple individuals. For some families, the data show that all three family members have similar responses on objective questions, such as how much time they spent together; as indicated by the high correlation between their responses. In these same families, however, the data exhibit much more disagreement, on subjective questions, such as how is the overall condition of the family, as indicated by low or none correlation.

4.2 Mobilize: Mobilizing for Innovative Computer Science Teaching and Learning

Mobilize (mobilizingcs.org) is an NSF-funded educational program that brings computational thinking, and data collection and analysis skills into the LAUSD STEM classrooms through the use of participatory sensing technology. During the deployment in 2013, 446 high school students from 21 classes in 8 schools used ohmage to submit surveys and interpret data on three different subjects—Explored Computer Science (ECS), Mathematics, and Science. In addition to using ohmage, some students also ran Systemsense, which is an Android-based system logging application that captures and analyzes participants' phone usage [13]. Lifestreams is used to answer the following two questions: 1) whether the students' engagement with the study varied across different classes or different subjects, and 2) whether the students' attachment to the phone impacts their data collection engagement. The first question is useful in understanding the effectiveness of different pedagogy methods, and the second question will help program managers decide whether to invest in phone access to encourage engagement.

The survey submission counts are used as a metric to measure the student's engagement in a class. The one-way ANOVA analysis module (see Section 2.4) reveals that the students' survey submission counts differed significantly across subjects, classes and teachers. For some teachers with multiple same-subject classes, the results show consistent pattern among their classes. These results indicate that classroom management including individual pedagogy method is a significant factor in student participation.

To answer the second question, we developed a data extraction module to extract phone usage features of the 90 students with the Systemsens data. These features include the interaction time with the phone, the amount of time the phone is power-on, the amount of network usage, the number of apps installed, camera usage events, and etc. A regression module is developed based on *lme4* package to estimate the effect of phone usage on the student engagement [41]. To adjust for the effects from different classes and subjects, we introduce two random-effect variables, one for class effects and the other for subject effects. The Class effect variable varies across different classes, but remains constant within the same class, and a similar property applies to the Subject effect variable [40]. Using this model, we found that logarithm of the mean phone interaction time and the amount of network usage both have small but statistically significant positive effects on the survey submission counts (coefficients = 0.082 and 0.045, both p-values < 0.0001). This indicates that the attachment to the phone may have effects on the students' engagement with the data collection. In total, 203 lines of code were added to implement the additional modules for this study, and 393 lines of code were reused.

5. LIMITATIONS AND CHALLENGES

Our analyses suffer from incomplete or missing data due to misunderstanding, participants' constraints, and technical problems such as phone malfunction and battery consumption issue. Currently, Lifestreams treat missing data in survey responses by excluding them from the analyses (e.g. pairwise deletion in the correlation analysis) and treat missing sensory data, such as activity, audio, and location data, by setting a cutoff threshold (e.g. any day that has less than 70% of coverage or less than 16.8 hours of data will be ignored). The missing data would cause problems in certain analyses such as the change detection. In the future, statistical methods of data augmentation will be used to infer and impute the missing data.

Furthermore, the phone-based sensing applications may sometimes fail to capture users' behaviors. For example, some participants did not carry their phones during exercise, and some left their phones stationary all day. While some of these specific usability problems can be mitigated by using wearable sensors such as Fitbit [42], and others were an artifact of the participant having a second primary phone, a broader challenge is to create a more engaging mobile user experience and more sophisticated data models.

Some statistical results (e.g. the number of changes detected per dataset) shown in the analysis are sensitive to the parameters. As with most statistical methods, many of the analysis techniques in Lifestreams require iteratively tuning and validation based on the "ground truth". In this paper, we use the qualitative interview data as the ground truth. However, participants' descriptions of their

condition might be inaccurate, or it might be hard or inappropriate to elicit certain information from the participants. Expert domain knowledge from behavioral and health science researchers is needed to provide further insight into and help validating the results.

This version of Lifestreams was designed for researchers or highly trained individuals who have familiarity with statistics or at least extensive experience in their domain. The analytic modules are intentionally highly configurable and provide detailed information that allows users to interpret the results. Such amounts of flexibility and information, however, would be overwhelming to users in a real-world clinical setting where time per patient is limited and regulated. Much work will be needed, both in analysis methods and interface designs, to summarize information and produce results that are more accessible to busy users. More generally, there is significant work to do related to the data collection, validation and generalization of these techniques.

6. RELATED WORK

Lifestreams utilizes contextual data analysis, data mining, and the collection of personal data streams. Therefore this related work section covers all three topics, starting with the use of mobile phones to collect personal data streams.

6.1 Collection of personal data streams

6.1.1 Active data capture with mobile phones

Traditional methods of self-monitoring through retrospective self-report are prone to errors and biases due to limitations of autobiographical memory [7, 36]. For this reasons, mobile devices offer the advantage of supporting active data capture from participants in the context of their everyday lives and closer to real time. A structured, in context, form of real time self-report, referred to in the literature as Ecological Momentary Assessment (EMA), was developed to monitor affect, cognitions, and behaviors in real time in a person's natural environment [21]. Error and bias are reduced by collecting and recording data in real time.

On the other hand, self-monitoring has been shown to encourage behavior changes and increases adherence by supporting self-awareness and self-efficacy [11, 12, 29]. Lifestreams is designed to supplement such interventions by making self-monitoring data more actionable by helping to identify patterns and trends in detailed datasets.

6.1.2 Passive data capture from mobile phones

In recent years smartphones have been increasingly used for passive data collection from on-board sensors. Phones can determine their location by detecting nearby cellular and Wi-Fi access points, or directly with GPS. Eagle and Pentland [13] conducted the Reality Mining study, gathering Bluetooth, GSM, and application usage data to build models for social systems. SystemSens and Funf [17, 3] likewise captured many data signals. Earlier, Ubifit, demonstrated the power of passive monitoring of activity for behavioral feedback on physical activity [10]. Lifestreams builds on this work, and is designed to integrate with personal data collection platforms, as we described in Section 2, through integration with the personal data collection platform ohmage.

6.2 Context data analysis

The data collected autonomously by mobile phones can be used to infer the smartphone's context, and by proxy, the user's as well. Activity classification and semantically meaningful locations are two such inferences that help phone to learn user patterns.

6.2.1 Activity classification

One important aspect of user context is activity classification using mobile sensors. Before smartphones became common, research in this area used specialized sensors [5]. More recently, with data from smartphone accelerometers and location sensors, phones can classify the user's activity, such as driving, walking, running, etc. [32, 35]. Features such as variance of magnitude of acceleration are calculated from the raw triaxial accelerometer data and, with GPS speed or other location change information, are used to infer the user's ambulatory or transport mode. These inferences lend themselves well to behavioral indicators relating to activity level.

6.2.2 Place detection

Another important type of context information relevant to users and their routines is location. Modern smartphones are equipped with GPS that ascertains geographic location given satellite reception. In addition, they also have Wi-Fi and cellular radios, which can be used to provide the user's rough location by identifying nearby access points [25].

More important than raw coordinates are semantically meaningful locations, such as home or work. Significant locations in a user's pattern can be learned through processing location traces to infer which locations, defined by a geographic location, are important to the user [4, 27].

Lifestreams analytical building blocks incorporate the ideas from this work. For example, location analysis and activity feature extraction building blocks process the raw sensor data and extract a stream of robust context values, which can be aggregated into higher-level features that are intensively used in the further analytic modules to enable powerful inference to participants' behaviors.

6.3 Data mining with mobile data

Several projects have demonstrated the power of data mining and machine learning to infer user-status from mobile phone data. The Reality Mining project [13] infers people's routines from their location (divided into home, work, elsewhere) from nearby cell towers and Bluetooth devices. BeWell [26] infers daily scores for sleep, activity, and sociality of users using various smartphone sensors. MoodSense [28] passively collects phone usage data such as emails, web history, and calls, as well as locations to estimate the user's mood. Eigenbehaviors [14], on the other hand, classifies human daily routines by using principle component analysis on temporal location data.

Lifestreams differs from these studies in two ways: (1) instead of providing several specific inference methods, Lifestreams is aimed to provide a suit of extensible analytics building blocks to facilitate the development of new inference functions. (2) As compared with some studies in which the inference results are used as a service or feedback to the user, Lifestreams' inference functions are more focused on providing researchers and designers of interventions with insights to the participants or patients' behaviors.

7. CONCLUSION / FUTURE WORK

The key contribution of this work is Lifestreams, which is a data analysis software stack consisting of four layers—*feature extraction and aggregation, feature selection, inference and interactive visualization*—as well as an initial set of analytical building blocks. Lifestreams facilitates the exploration and evaluation of multi-dimensional personal data streams for potential behavioral indicators.

Using Lifestreams' analytical pipeline, we present three sets of analysis results based on our inference methods to help make sense of a real-world dataset. These inference methods identified significant behavioral trends and patterns, and their accuracy were qualitatively verified in follow-up interviews conducted with participants. In addition, Lifestreams was found to be useful as a tool for the research coordinator to quickly navigate through the data and provide visual aids to guide the discussion with participants during interview sessions.

Lifestreams is designed to be an extensible system that can be generalized and applied to different studies containing varying data streams. We evaluated Lifestreams' extensibility and generalizability by applying it to two additional studies with unique requirements and new types of data streams.

Ultimately, researchers will develop interventions that can greatly improve the personalization and precision of patient-centric care. For example, consider a rehabilitation program for patients with hip surgery. A healthcare provider could deploy a mHealth study in which 50 patients are asked to use phones to passively monitor their mobility. During the study, the provider would use a tailored instance of Lifestreams to check the data weekly and identify if patients are having any trouble complying with the study requirements. After a few months, when a large amount of data has been collected and is ready for the provider to analyze, she could use Lifestreams to study the interaction between patients' recovery progress and their daily life behaviors; such as working schedules, sleep patterns, stress and mood, each automatically extracted by the Lifestreams' feature extraction modules. Moreover, the provider could use Lifestreams to monitor if significant change occurs in the patients' recovery progress and other related behaviors in response to a new treatment plan. Assisted with the visualizations generated by Lifestreams, the provider could use this information to guide the discussion with patients and families, to provide feedback, to identify triggers of the changes, and to help patients with problem solving, goal setting, and monitoring progress towards goals.

Personal data stream platforms collect a wealth of high-resolution data about an individual and their context. However, the promise of mHealth to truly revolutionize patient care and actualize the information collected lays in the ability to translate the information collected into simple and meaningful insights about an individual. Lifestreams attempts to take the first step to achieve that goal.

8. ACKNOWLEDGMENTS

This work was made possible with funding from UCLA Center for Embedded Networked Sensing (NSF #CCF-0120778), the National Science Foundation Math Science Partnership (NSF #DUE-0962919), the National Institutes of Health (NHLBI #5RC1HL099556), the Intel Science and Technology Center for Pervasive Computing (ISTC-PC), Cornell NYC Tech, and gifts from Google and Nokia. The authors would like to thank Steve Nolen for his support of the deployments; Hossein Falaki and Dony George for their data collection tools; and anonymous reviewers for their valuable suggestions which that have helped improve the quality of this manuscript.

9. REFERENCES

[1] The free dictionary: Inference. http://www.thefreedictionary.com/inference.

[2] Ginger.io - http://ginger.io/.

[3] N. Aaron. A data-rich approach for investigating social mechanisms in the wild. In *Ubicomp'11*.

[4] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[5] L. Bao and S.S. Intille. Activity recognition from user-annotated acceleration data. *Lecture Notes in Computer Science*, pages 1–17, 2004.

[6] M Basseville and IV Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, NJ, 1993.

[7] N.M. Bradburn. Temporal representation and event dating. *The science of self-report: Implications for research and practice*, pages 49–61, 2000.

[8] P. J. Brockwell. *Time Series Analysis*. John Wiley Sons, Ltd, 2005.

[9] C. Zhou, D.Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personal gazetteers: an interactive clustering approach. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*.

[10] S. Consolvo, D.W. McDonald, T. Toscos, M.Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. SIGCHI'08.

[11] M.L. Dansinger, J.A. Gleason, J.L. Griffith, H.P. Selker, and E.J. Schaefer. Comparison of the atkins, ornish, weight watchers, and zone diets for weight loss and heart disease risk reduction. *JAMA: the journal of the American Medical Association*, 293(1):43–53, 2005.

[12] D.M. Donovan and G.A. Marlatt. *Assessment of addictive behaviors*. Guilford Press, 2005.

[13] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[14] N. Eagle and A.S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–10, 2009.

[15] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96*.

[16] D. Estrin and I. Sim. Open mhealth architecture: an engine for health care innovation. *Science*, 330(6005):759–760, 2010.

[17] H. Falaki, R. Mahajan, and D. Estrin. Systemsens: a tool for monitoring usage in smartphone research deployments. In *MobiArch'11*.

[18] Bailey, Rosemary A. Design of comparative experiments. Vol. 25. Cambridge University Press, 2008.

[19] T. Fu, F. Chung, V. Ng, and R. Luk. Evolutionary segmentation of financial time series into subsequences. *Evolutionary Computation, 2001*, 1:426–430, 2001.

[20] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes. Learning and recognizing the places we go. *UbiComp'05*.

[21] M.R. Hufford. Special methodological challenges and opportunities in ecological momentary assessment. *The science of real-time data capture: Self-reports in health research*, pages 54–75, 2007.

[22] D.H. Kim, J. Hightower, R. Govindan, and D. Estrin. Discovering semantically meaningful places from pervasive rf-beacons. In *Ubicomp'09*.

[23] P. Klasnja and W. Pratt. Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, 45(1):184 – 198, 2012.

[24] T.L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society*, 57(4):613–658, 1995.

[25] A. LaMarca, et al. Place lab: Device positioning using radio beacons in the wild. *Pervasive Computing*, pages 301–306, 2005.

[26] N.D. Lane, M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury, and A.T. Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *PervasiveHealth'11*.

[27] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.

[28] R. LiKamWa, Y. Liu, N.D. Lane, and L. Zhong. Can your smartphone infer your mood. In *PhoneSense'11*.

[29] G.A. Marlatt. Relapse prevention: Theoretical rationale and overview of the model. *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors*, pages 3–70, 1985.

[30] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[31] Cohen, Jack. Statistical power analysis for the behavioral sciences. Routledge Academic, 1988.

[32] S. Reddy, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Determining transportation mode on mobile phones. In *Proceedings of The 12th IEEE Int. Symposium on Wearable Computers*, 2008.

[33] Gordon J. Ross. cpm: Sequential Parametric and Nonparametric Change Detection, 2012.

[34] J. Ryder, B. Longstaff, S. Reddy, and D. Estrin. Ambulation: a tool for monitoring mobility patterns over time using mobile phones. In *2009 International Conference on Computational Science and Engineering*, pages 927–931, 2009.

[35] T. Saponas, J. Lester, J. Froehlich, J. Fogarty, and J. Landay. iLearn on the iPhone: Real-time human activity classification on commodity mobile phones. *University of Washington CSE Tech Report UW-CSE-08-04-02*, 2008.

[36] N. Schwarz. Retrospective and concurrent self-reports: The rationale for real-time data capture. *The science of real-time data capture: Self-reports in health research*, pages 11–26, 2007.

[37] D. Wied and P. Galeano. Monitoring correlation change in a sequence of random variables. *Journal of Statistical Planning and Inference*, 143:186–196, 2013.

[38] C. Zhou, C. Zou, Y Zhang, and Z Wang. Nonparametric control chart based on change-point model. *Statistical Papers*, 50(1):13–28, March 2007.

[39] R. Fenwick, and M. Tausig. Scheduling stress family and health outcomes of shift work and schedule control. *American Behavioral Scientist* 44.7 (2001): 1179-1198.

[40] H., Donald, Generalized linear mixed models. *Encyclopedia of statistics in behavioral science*, 2005.

[41] *lme4* - Mixed-effects models project. http://lme4.r-forge.r-project.org/.

[42] Fitbit - http://www.fitbit.com/.

[43] B. Longstaff, S. Reddy, D. Estrin. Improving activity classification for health applications on mobile devices using active and semi-supervised learning, *PervasiveHealth*, 2010.

[44] Passive PRO - http://c3nproject.org/innovations/patient-self-tracking-continuous-care/passive-pro

[45] Mobile App: PTSD Coach. National Center for PTSD - http://www.ptsd.va.gov/public/pages/ptsdcoach.asp

[46] C. C. Quinn, S. S. Clough, J. M. Minor, D. Lender, M. C. Okafor, and A. Gruber-Baldini. WellDoc mobile diabetes management randomized controlled trial: Change in clinical and behavioral outcomes and patient and physician satisfaction. *Diabetes Technology and Therapeutics*, 10(3):160–168, 2008

APPENDIX A: Ohmage: An Extensible, Open Source Software Platform

In this appendix, we briefly introduce *ohmage*, which is a personal data collection platform Lifestreams is built on top of. Ohmage (http://ohmage.org) is an open-source, mobile to web platform that supports the collection, storage, analysis and visualization of personal data streams. These include intermittent self-report data entered by the user, as well as continuous streams of data passively collected from sensors or applications on-board the mobile device (e.g. activity and location traces). Ohmage additionally includes rich system and user analytics to instrument the act of participation itself and ultimately to provide user contexts, patterns and interventions. Ohmage is feature-rich and extensible, and facilitates the collection of multi-dimensional, heterogeneous and complex personal data streams. Ohmage is therefore an ideal platform for integrating with Lifestreams for development and evaluation.

Ohmage is divided into three primary components: 1) the ohmage server (which includes a private back-end datastore), 2) the ohmage client app and supporting phone software, and 3) the ohmage administrative website.

The phone client also acts as an aggregator through which other ohmage plugins can upload sensor data for the user to ohmage. Rather than being defined with the study, sensor data that is collected continuously in the background (also called 'passive data collection') is instead defined globally in an ohmage's sensor data framework and collected for a participant outside the context of a particular study. This approach avoids duplicating the large volume of sensor data while still allowing it to be collected and analyzed alongside the active, study-specific data.