Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change

FRANK BENTLEY, Yahoo! Labs[1]
KONRAD TOLLMAR, Royal Institute of Technology
PETER STEPHENSON, Humana
LAURA LEVY, Georgia Institute of Technology
BRIAN JONES, Georgia Institute of Technology
SCOTT ROBERTSON, Georgia Institute of Technology
ED PRICE, Georgia Institute of Technology
RICHARD CATRAMBONE, Georgia Institute of Technology
JEFF WILSON, Georgia Institute of Technology

People now have access to many sources of data about their health and wellbeing. Yet, most people cannot wade through all of this data to answer basic questions about their long-term wellbeing: Do I gain weight when I have busy days? Do I walk more when I work in the city? Do I sleep better on nights after I work out?

We built the Health Mashups system to identify connections that are significant over time between weight, sleep, step count, calendar data, location, weather, pain, food intake, and mood. These significant observations are displayed in a mobile application using natural language, e.g. "You are happier on days when you sleep more." We performed a pilot study, made improvements to the system, and then conducted a 90-day trial with 60 diverse participants, learning that interactions between wellbeing and context are highly individual and that our system supported an increased self-understanding that lead to focused behavior changes.

Categories and Subject Descriptors: CCS → Human-centered computing → Ubiquitous and mobile computing → Empirical studies in ubiquitous and mobile computing; CCS → Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → Field studies; CCS → Applied computing → Life and medical sciences → Health informatics

General Terms: Design, Experimentation, Human Factors

Additional Key Words and Phrases: Health, Context, Mobile, Wellbeing, Mashups

ACM Reference Format:

Bentley, F., Tollmar, K., Stephenson, P., Levy, L., Jones, B., Robertson, S., Price, E., Catrambone, R., Wilson, J. 2013. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change.

DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The Merriam-Webster Dictionary defines "wellbeing" as "the state of being happy, healthy, or prosperous." [Merriam-Webster, 2011] Wellbeing is an important measure of the quality of a person's life and is a measure that we, as social computer science researchers, hope to increase through new computational systems that help people understand the patterns in their lives that impact their wellbeing over time.

Much of the world today is experiencing a breakdown of general wellbeing at a scale never before seen. In America, over one-third of adults are officially obese as are 17% of children. [CDC, 2010] Western-European countries are also quickly approaching these numbers. For example, one in six children aged between two and 15 are obese in the UK. [BUPA, 2011] Major factors contributing to obesity include a sedentary lifestyle and high-calorie food choices. Many aspects of people's lives can lead to these lifestyle choices. Some people live in cultures of fast and fried foods, rich desserts, and peer pressure to eat to excess. [Christakis and Fowler, 2007] For others, a busy or car-centered lifestyle can lead to less time for physical activity. This sedentary lifestyle leads to an increased risk of chronic diseases, the leading cause of death in the world, including hypertension, diabetes and obesity.

Another challenge for wellbeing is sleep. Millions of people around the world have trouble sleeping, many to the point where it severely impacts their ability to function during the day. It is

[1] Work conducted while at Motorola Applied Research Center

often quite difficult to understand what factors lead to a better night's sleep. Many millions also suffer from mood disorders, and understanding the aspects of life that affect mood is often a critical, but difficult, first step towards improvement.

A variety of contextual factors can work together to impact a person's overall wellbeing. The weather, food, alcohol or caffeine intake, stress levels, or having late-night or early morning appointments all can affect sleep as well as daily activity levels and weight. We are interested in understanding the effects of these contextual variables on wellbeing over time and how people change their behavior with increased awareness. All of these aspects of wellbeing are able to be captured fairly easily using devices on the market. However solutions that combine data from all of these sources are not yet prevalent, much less ones that look for longer-term trends.

Therefore, we set out to study these long-term patterns in individuals' daily lives in order to make people aware of them and reflect upon. Managing an overall state of wellbeing is a game of tradeoffs. A slice of cake may contribute to your happiness today, but not to your health in the longer term. Making one bad decision of this magnitude will not impact overall wellbeing, but over time such decisions add up to create patterns and long-term effects. [Rachlin, 2004] We will argue that by providing a tool to help people understand long-term patterns of behavior, users will be empowered to see the tradeoffs that they face in daily life in new ways that are difficult to spot on their own. Understanding these trends helps to focus behavior change to specific and actionable moments. For example, discovering trends in personal information indicating that on days when you have many scheduled meetings you gain weight could cause you to reflect on why that is the case and take action to change it on your next busy day by opting for the salad over the French fries in the cafeteria or explicitly reserving time for exercise. We initially postulated that these correlations between context and aspects of wellbeing would be quite different for each individual. Therefore, we have built a system that analyzes an individual's wellbeing across multiple dimensions and identifies the significant patterns that emerge over time with respect to a variety of automatically-gathered contextual data.

It is currently quite difficult for people to discover these long-term patterns about themselves. Even for those who use tools like the Withings scale and Fitbit pedometer to track their daily weight and step count, it is not possible to see trends between the two or to see how they interact on specific days of the week, weekends vs. weekdays, month to month, etc. over time without exporting the data into complex statistical packages [Li et al, 2011]. Users are facing a data overload in which the complex interactions among streams of weight, steps, workout data, calorie intake data, location, calendar information, and more are just too much to process on one's own, especially through time-series graphs. Our work is based on the idea that that seeing cross-sensor insights into one's wellbeing over time is needed to create an understanding of the underlying causes of negative behaviors and to start on a path towards contemplative action. [Prochaska and Velicer, 1997]

Smartphones are a useful platform for collecting contextual data from many aspects of a person's life and for discovering trends that illuminate correlations between aspects of a person's context (e.g. total time free/busy in a day, location, weather, etc.) and their wellbeing (e.g. weight change, daily step count, sleep, mood, etc.). People carry smartphones throughout the day and use them frequently each day in small bursts of time. By displaying significant observations about a person's wellbeing on the phone, we aimed to encourage reflection on wellbeing in these small breaks, much along the lines of previous work with step counts (e.g. [Connelly et al, 2006; Consolvo et al, 2006, 2008, 2009]).

This paper will discuss our system, the first individually-focused platform for automatically finding significant trends in long-term wellness and context data, as well as findings from two field studies, an initial pilot study and a larger 60-participant, 90-day field study with an improved version of the system. We will discuss how participants were able to reflect on their wellbeing in new ways through this mobile-based system and how they created focused behavior change strategies based on the observations that were presented to them in the system. These behavior changes led to significant improvements in weight, mood, and WHO-5 wellbeing scores over the course of the study. This work can serve to help others in creating wellbeing systems that cause

users to reflect on the impact of patterns in their daily lives and make positive changes to their behavior to improve mood, weight, activity levels, sleep, or other aspects of their general wellbeing.

2. BACKGROUND AND DESIGN MOTIVATION

Over the past five years, there has been a solid stream of work on mobile wellbeing systems both in the HCI community and in the commercial applications/devices space. Commercially, devices such as the Fitbit and Nike+ sensors have allowed people to examine their physical activity at a great level of detail. Similar devices such as Philips Direct Life provide easy ways to understand daily activity levels and provide simple suggestions on ways to be more active throughout the day. Internet-connected scales (e.g. the popular Withings model) allow people to easily keep track of their weight and changes over time without the need for manual log-keeping.

While some of these services, such as Fitbit, allow users to import data from multiple sensors (e.g. Fitbit and Withings) into a single account, these commercial services currently do not provide any graphs, insights, or suggestions to users based on the combination of different wellbeing data feeds. Each sensor is devoted to its own space in the interface. For example, graphs on the Fitbit website show information regarding the number of steps the user has walked in one box and a graph of the user's weight in another with no way to directly compare them or to easily discern patterns in the data over time.

In the research community, Li et al have developed a system to display contextual information with related wellbeing data in time-series graphs [Li et al, 2011]. This system allows users to remember the context surrounding specific single scenarios when trying to interpret spikes or valleys in data such as step count or weight. This work shows the importance of understanding the connections between context and wellbeing data, but only showed the potential on a small scale of what happened on a particular day. We believe that this is an important start, but that long term trends and correlations are still quite difficult to discover in such systems. This is why we focus on mining wellbeing and contextual data streams for correlations and deviations over months of data. In his PhD thesis [Li, 2011], Li agrees and says that correlation analysis of this data is "difficult" and "another research project."

Other research systems such as Salud! [Medynskiy and Mynatt, 2010] also aggregate data from multiple sensor and contextual streams. In Salud!, as in Li's work, users are left to interpret the data on their own, from time series graphs and large data tables. It can be quite hard for users, especially those not able to interpret graphs, to gain any meaningful insights from this data. Hence, we are interested in automatically analyzing similar data sets using statistical methods to identify long-term patterns for users.

Studies of numeracy and statistical literacy have shown that many lack the ability to understand and apply data from graphs. Galesic and Garcia-Retamero [2011] found that 41% of Americans as well as 44% of Germans had low graph literacy skills in understanding very simple bar and pie charts. Ancker and Kaufman [2007] take a broader look at health numeracy and discuss not only problems in interpreting graphs, but also problems in understanding statistical data. These studies motivated us to consider alternate methods to present the complex interactions between wellbeing data streams over time and between sensors.

While there appear to be no clear practices on how to verbally convey statistical data [Gigerenzer et al, 2007], several suggestions have emerged from the literature. Lipkus [2007] discusses using variations on the phrase "likely" to discuss risk or statistical certainty. We chose to adopt this in appending "very likely" on the end of any observation with high statistical certainty ($p < 0.01$). For the rest of the formulation of the natural language sentences that we presented to the user, we were largely on our own. We debated several formats and settled on forms that were quite neutral and did not convey a particular need for action, leaving this to the user's interpretation. This resulted in sentences of the form "On days when you X, you Y" or "On Wednesdays you X more than usual."

Consolvo et al have explored mobile systems to encourage people to be more active in their daily lives. They have built and field-tested several prototypes in this domain starting with Houston, a

system to track step counts and share them with friends or family to create a competition/game around being active in daily life. [Consolvo et al, 2006] In another system, they used the mobile homescreen to display physical activity logged by UbiFit [Consolvo et al, 2008], a system that allows people to visualize their individual physical activity in the form of a garden that grows on the homescreen of the phone as a user performs a wide set of physical activities. They showed that users with the awareness display were able to better maintain their level of physical activity compared with those who did not use the display. These systems showed the promise of the mobile platform for wellbeing related behavior change, even while focusing only on physical activity.

Consolvo et al have also developed a set of design guidelines for systems that support health behavior change [Consolvo et al 2009]. While many of these guidelines focus on systems for people already making change, some are relevant to consider for systems that are focused on initiating change, especially those around being reflective, positive, controllable, historical, and comprehensive. We focused on these guidelines when creating our system.

BJ Fogg has created a series of guidelines for behavior change [Fogg, 2002, 2003] and has been exploring the mobile device as a platform to encourage behavior change. [Fogg and Allen 2009] He has explored text messaging as a means to encourage behavior change and the power of triggers in notifying users of potential change opportunities in addition to learning about their current progress. [Fogg and Allen 2009]

Anderson et al developed a system called Shakra [Anderson et al, 2007] in which users' physical activity is monitored using the GSM cell signal information. In their study, they found that making the user aware of their daily activity encouraged reflection and increased motivation for achieving high activity levels. We were encouraged by this finding and also by the possibility of using the mobile phone as a sensor of environmental context.

Based on the above work and the availability of consumer-grade sensors to detect various streams of data, we chose to focus on step count, sleep data, food, exercise, calendar data, and location for our first pilot study. These captured a wide range of wellbeing sensors and contextual information and would be sufficient to validate our hypotheses on finding significant patterns in personal wellbeing data and context. The overarching goal was to have no sensor stream take more than a few seconds of interaction per day to keep the burden of using the system as low as possible over time. Other, harder to sense, data could of course be added in the future such as stress levels, caloric or salt intake, etc. The research on these types of data streams indicated that they were just not ready for wider deployment in sustained daily use due to the complexity of manual entry [Connolley et al, 2006] or the burden of capture and analysis [van Eck et al, 1996].

Lastly, there exists a vast amount of related work on behavior change in social psychology and preventive medicine. For example, Emmons & McCullough [2003] demonstrated in an experimental study that people who kept weekly "gratitude journals" exercised more regularly, reported fewer physical symptoms, felt better about their lives as a whole, and were more optimistic about the upcoming week compared to those who recorded hassles or neutral life events. Kahn et al. [2002] reviewed and evaluated the effectiveness of various approaches to increasing physical activity in preventive medicine. Besides direct "point-of-decision" prompts to encourage physical activity they describe approaches such as behavioral and social interventions, individually adapted health behavior change, and environmental and policy interventions. The review summarizes the actual effectiveness of these approaches but also highlights the complex dependencies among various information interventions. Multiple sources of information need to come together effectively to create positive changes and this is what we are trying to achieve with our system.

Along the lines of Mamykina et al [2008], we sought to support reflection on personal data, and through this reflection our users could be better prepared to make lasting changes to their behavior. However, as Mamykina's study showed, individuals are often not comfortable with analyzing their own data. They want physicians or other professionals to reach the conclusions for them. We saw the text-based observations in our system as a good middle ground for reflection that did not require interpretation of the raw data by the individual user.

3. PILOT SYSTEM

To start exploring the concept of personalized health mashups, we created a pilot system that we fielded with ten participants to examine whether the system could produce useful observations with the types of data that people can log about themselves. Further, we wished to see if these observations could lead to behavior change. Findings from this pilot study were then used in constructing the complete system that was trialed with 60 participants for 90 days. The results of this larger study will be discussed in Sections 5 and 6.

The pilot system consisted of a Mashup Server that interfaced with the contextual and wellbeing data for each user from multiple data sources, performed a statistical analysis across the data, and presented the user-specific observations as a natural language feed to a mobile phone widget (as shown in Figure 1). This widget then linked to additional graphs and data that users could explore to dig deeper into the details of their wellbeing.

	Pilot Study	Full Study
Automatically Sensed		
Location (City-level)	**X**	**X**
Weather		**X**
Calendar Free/Busy Hours	**X**	**X**
Sensor Inputs		
Step Count (Fitbit)	**X**	**X**
Sleep (Fitbit)	**X**	**X**
Weight (Withings scale)	**X**	**X**
Manually Logged		
Food	**X**	**X**
Exercise	**X**	
Mood		**X**
Pain		**X**

Table 1: Data streams that were used in the Pilot study and the full study.

We collected data from several sources as shown in Table 1. Data sources connect to the open APIs of each commercial sensor and use custom REST APIs for data coming from our own context logging software on the phone. Data from the phone included automatically capturing hours "busy" per day from the calendar, location at a city level, and manual logging of daily food intake and exercise.

Each night, we performed a statistical analysis of the data for each user and updated the feed of significant observations per user. The feed could include observations across data types such as "On days when you sleep more you get more exercise" or observations from a particular sensor over time: "You walk significantly more on Fridays." For the first type of correlation, we surfaced significant observations based on statistically significant Pearson correlations between sensors ($p < 0.05$). For the deviations, we surfaced differences greater than a standard deviation from the mean. For day of week differences, we performed a t-test and displayed significant results ($p < 0.05$) in the feed. We performed the analysis based on different time scales: daily, weekly, monthly and by day of the week. As such, it could include observations such as "Last week you slept significantly less than usual." Additional technical detail on the statistical methods used can be found in [Tollmar et al 2012].

While Pearson correlations are relatively simple, we believed that they would be an understandable and meaningful measure of how aspects of a person's life were related. Along with the deviations between days and for individual days, they also could easily translate into a natural language sentence. Other statistical methods, e.g. the intra-day modeling in Albers et al [2010], could certainly be applied, however we wanted to start with simple techniques that had the highest chances of being understood in natural language. By performing our analysis over weeks and months, in addition to daily, we were able to find some effects that had some delay, such as weeks with higher step counts leading to weight loss for some participants where this effect would not be visible on a daily basis.

Data from the feed was presented in a widget on each user's phone. Only statistically significant observations were displayed in the mobile widget and all items contained a plain text confidence (e.g. "possibly," "very likely," etc.) based on the corresponding p-value. The content on this widget was updated each evening after the server computed the newly significant observations. The widget can be seen in on the bottom-left in Figure 1.

Clicking on a feed item in the widget launched our mobile website on which users could view a graph detailing that particular item. From this mobile website, users could also navigate to other graphs such as a day-by-day weight graph, average step count by day of week, other correlations to elements of their context, or to other streams of wellbeing data. If desired, more technical users could spend time exploring these graphs in order to better understand the data behind the observations. Additionally, users could mark observations to reflect on later in a favorites list accessible from the mobile website.

Beyond the mobile website, users could visit the website from their computers, where a larger view of the feed was visible and it was a bit easier to navigate between graphs and data from different sensors.

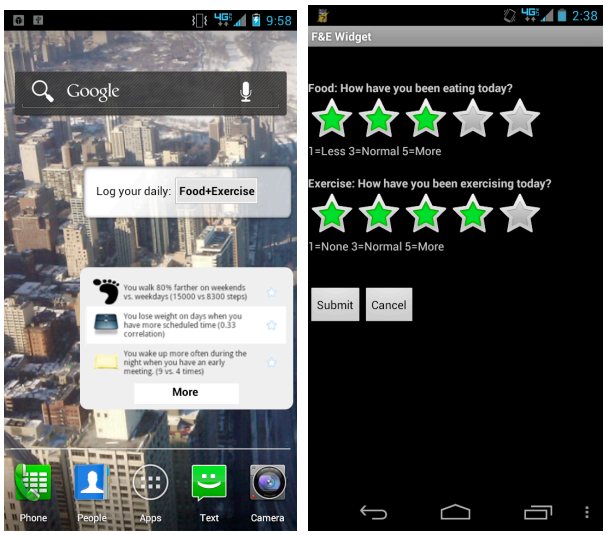

Figure 1: The pilot system provided mashups observations in a widget on the homescreen as well as on a mobile website. Manual logging of food and exercise was accomplished via another widget, which opened the logging screen on the right.

On the phone, we developed a simple application for reporting general food and exercise behaviors for the day as shown on the right in Figure 1. Users were presented with a 5 star scale and asked to rate their food intake and exercise for the day. We suggested that participants log this activity at the end of each day as a way to reflect on the day's behavior. While we are aware of the problems with users often not supplying regular and sustained data with manual logging [Burke et al, 2008; Patrick et al, 2009], we felt that these components were important to provide for users who were motivated to log and understand these elements of their lives and that the simple, 2-click logging would be a smaller burden than in the previous studies.

The system was intended to be used daily, where users step on a scale, wear the Fitbit, have their phones upload contextual data in the background, and log food and exercise habits each evening. These are all fairly low-effort activities that we hoped could easily become a part of our participants' daily lives. The mobile phone widget allowed users to be frequently reminded of what is most significantly affecting aspects of their wellbeing whenever they looked at their device. We speculated that this system would then encourage positive behavior changes based on personal reflection on this data, and designed a pilot study to investigate how this system would be used over time in the daily lives of participants in two major cities and to learn how the concept could be improved for a larger study.

4. PILOT STUDY

4.1 Methods

We recruited ten diverse participants for a two-month field evaluation of the pilot system in the summer of 2011. Four users lived in Chicago (from here on referred to as C1-C4) and six were from Stockholm (S1-S6). The Chicago participants were recruited through a professional recruiting agency to reflect a diversity of age, occupations (e.g. police officer, real estate agent, waitress, chemist, etc.), cultural backgrounds, and education. The Stockholm participants were recruited through extended social networks of the research team and also included a variety of ages and occupations.

The study lasted for two months. We began the study with a semi-structured interview in the participants' homes. We explored their current wellbeing practices as well as any goals they might have around their health or general wellness. We helped the participants set up the Wi-Fi scale and Fitbit devices and instructed them on their use. Participants also completed a demographic survey.

For the first four weeks of the study, participants used the scale and Fitbit in their daily lives and could use the websites provided by the device manufacturers to follow their progress. However, they did not have our mobile application installed or access to the observation feed generated by the Mashup server. We did this in order to get an understanding of what people can infer from the existing solutions without seeing the mashed up wellbeing data as well as to get a few weeks of background data from which to create initial observations for our feed in the second part of the trial. During these first four weeks, participants were instructed to call a voice mail system or send an email whenever they had an insight about their own wellbeing. We were interested if they would make any insights that involved multiple sensors or patterns over specific days of the week, but did not inform participants of this fact, or what the mobile app in the second month would entail. We saw this first month as our "control" as we could see how participants would use existing tools in their daily life and what they could infer from the data presented in these services.

After the first month, we met with the participants again in their homes and asked follow up questions based on their voice mail messages and any additional insights they were able to make. At this time, we installed our mobile widget/app on their phones and demonstrated its use. We also installed the contextual logging services including calendar free/busy data upload, city-level location sensing, and the manual food/exercise logging widget. Participants had the ability to turn off the background contextual logging at any time and some participants did not provide data for all context attributes (e.g. if they did not have their Android calendar populated). For the final four weeks of the study, participants were free to continue using the websites from the individual sensors but also had access to our widget, mobile website, and full website to further explore the connections between the different aspects of their wellbeing. They were also instructed to continue to call into the voicemail system or send an email whenever they had a new insight about their wellbeing. Participants were told to use the system as if they just downloaded the app from the market and that no compensation would be tied to their use (or non-use) of the system.

At the end of the second month, we conducted final interviews (over the phone in Chicago and in-person in Stockholm) reviewing insights that participants were able to make and general comments on the system itself. Participants were able to keep the scale and Fitbit as a thank you for participating in the study and were able to keep the widget on their phone if they wanted. Most participants chose to keep the widget.

In addition to the qualitative data from the interviews, voicemails and emails, we logged accesses to the web site (both desktop and mobile) to better understand the use of the graphs and feeds of significant observations. In addition, we also have a log of all data uploaded from the sensors themselves including weight, steps per day, hours sleeping, times awoken during the night, city-level location per day, calendar free/busy data per day, and any manually logged food and exercise data. Unfortunately, as the widget was on the home screen of the device, we have no way of knowing how many times each user looked at the widget if they did not click through to the mobile website for additional information.

After the study, all qualitative data was analyzed using a grounded theory based affinity. The items of analysis represented exact quotes from users (or English translations for Swedish participants) with groups and themes formed based on these direct statements. All themes discussed below have support from multiple study participants. Quantitative data from usage logs and the demographic survey were also analyzed using statistical tools and will be discussed in more detail below.

4.2 Findings

4.2.1 Use of the system

From the website logs and initial survey data, we performed a quantitative analysis to better understand the use of the system. There was a significant difference in the amount of use of the website between the two countries. Over the second month, when they had access to the application, participants in Sweden accessed an average of 70 pages while participants in Chicago only accessed an average of only 10 pages ($t(8) = 3.0$, $p < 0.03$). Participants in Sweden also walked more than twice as much each day compared to participants in Chicago (10792 steps vs. 5147 steps, $t(8) = 3.5$, $p < 0.01$).

Other wellbeing data did not differ significantly between the two locations. Starting weight, total weight lost/gained, variance in daily weights, and variance in weekly step counts did not significantly differ between the locations.

Eight of the 10 participants lost some amount of weight during the study, averaging 1.6kg. There were no significant differences by country, gender, or starting weight. This data is interesting because the Chicago and Stockholm groups used the website quite differently but had similar outcomes in terms of weight loss. If we had more complete step count and sleep data (see Section 4.2.3), it would also be interesting to analyze improvements in these aspects of wellness across cities.

Overall, the use was less than expected, especially the frequency of manual logging as described in Section 4.2.3. Therefore, we began to develop ways to improve user engagement with the system so that the quality of the observations could be increased. These ideas will be presented as design recommendations in the following sections and the final implementation of these recommendations will be discussed in Section 5 on the full system.

4.2.2 Learning from observations

Despite the limited data that was provided to create the observations, participants found certain entries in their feeds to be interesting and they were able to learn more about themselves through the widget and mobile website. User S5 was able to piece together two observations that told her that when she eats more she sleeps more but also that when she sleeps more she exercises more. Thus for her, eating more (i.e. enough) could lead to healthier sleep and thus a desire to be more active the next day and feel better overall. She found this quite interesting.

The day of the week observations were also quite useful to better understand trends over time. Participant C1 understood a feed item that told her that she tends to gain weight on Mondays: “It’s absolutely true! Cause on the weekends, like last Sunday, I went to my mom’s house and she made blackberry cobbler and I ate some of it.” C2 liked that the correlations were made explicit in the daily feed: “Like if you eat too much or booze too much your weight is going to go up. But it’s one thing to know it and another to actually see the results of what you did.”

Many individual correlations made sense to users such as C3 who said: “I eat less on days when I walk more, which I think is interesting. I think it means that just when I’m more active I tend to eat less and I think it’s probably accurate because when I eat more it’s probably because I’m bored or snacking.” She mentioned this correlation several other times during the study and it was clear to us that this was something she reflected on periodically, likely due to it appearing as a significant correlation for much of the study.

These insights from the observations came in contrast to insights reported during the first half of the study when participants did not have the mashup data available. Before getting the widget,

participants had insights about particular entries (e.g. a lower step count than expected on a certain day) or about the time series of one sensor (e.g. not losing weight as quickly as they had hoped they would). However, no participants had insights across sensors, days of the week over time, or with other aspects of their context, such as the ones enabled by the system in the second half of the study. Therefore, we see the system as a success in providing new and deeper ways for people to understand themselves and what contributes to their wellbeing.

The "scientific" nature of the system also appealed to participants. C1 liked that the system is "very truthful. It doesn't hide or anything like that. … It makes me know that I'm not reaching it and that I need to be doing physical activity." S2 also liked the "objective data" and having "metrics" about his wellbeing. Overall, most participants found interesting correlations and enjoyed this new way to learn more about their own wellbeing and appreciated the easy-to-understand, natural language presentation. We saw this as quite positive for the concept overall, if we could address the large issues of engagement in the next iteration.

4.2.3 Lack of Data Richness

In order for the system to operate well and provide accurate observations to our users, it is necessary to have as much data as possible from multiple inputs on the same day. We then need examples of good, bad, and average days in order to find patterns. However, many of our study participants did not use the devices with this regularity, especially the manual logging, making the overall feeds less detailed and reliable.

This sparseness of data was shown for several sensors as shown in Figure 2. Participants, especially those in Chicago, did not always wear their Fitbits all day and sometimes just took them out for times of exercise. Across all participants, manual logging did not occur frequently. Food and Exercise were logged on average only 5 times by each participant over the 30 days of the second half of the study. This is consistent with previous work on self-logging [Burke et al, 2008; Patrick et al, 2009] and a main focus of the design improvements for the full study would be on improving this number.

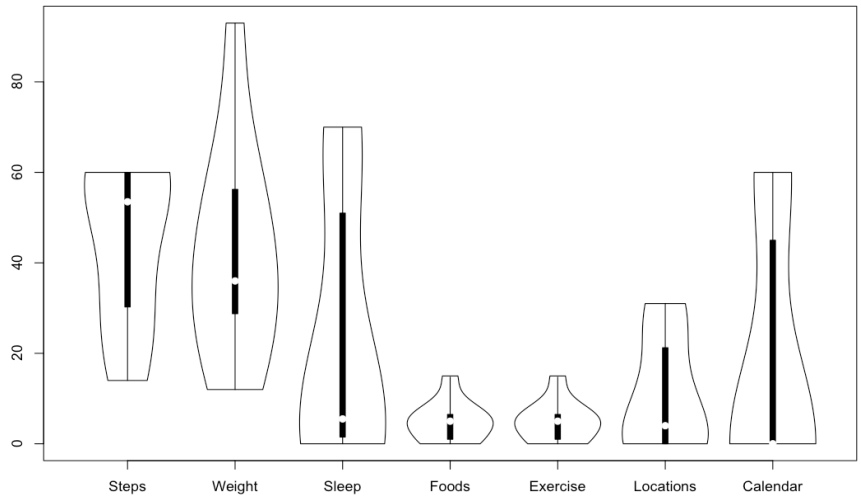

Figure 2: Average number of data points recorded for each sensor in the pilot study. The study lasted approximately 60 days (with Food, Exercise, and Location logging only the final 30 days). Note that some users took their weight or recorded sleep multiple times per day. Manually-entered streams of food and exercise were not frequently logged, and this would become a focus for our design iteration.

At times, some participants forgot about the widget or the sensors. Most of these problems were due to these users placing the widget on a secondary home screen that was not often displayed due to the widget's size and their already customized home screen setups. Thus, they needed to swipe the screen to find the widget and it was sometimes forgotten. As C2 said, "I'm never over on that screen. But if it was smaller then it could be on the main page or the next main page." One way to deal with this issue could have been to build a smaller widget with fewer observations that could more easily fit with the other content on the main screen.

Because of the lack of data provided, we often could not calculate specific correlations for users as only a small handful of days had readings from particular combinations of sensors. We realized that in order to have a system that produced meaningful results, we would need to have a way to engage users to provide more data over time.

4.2.4 Contradictory Information

Because of the lack of consistent use of each of the sensors, at times the system provided contradictory information over the course of a week. A correlation that might have been positive one day could swing negative with a few extreme data points on the other side in the following days. Our original goal was to use the four weeks of data from the first half of the study to ensure that the historical data would be more complete by the time of the second month. However, for several participants some data was only present for a few days (e.g. sleep) or only on particular days of the week making daily correlations or deviations based on days of the week quite variable with the addition of new data.

In his analysis of persuasive systems, Fogg warns that systems that produce questionable data are likely not to be trusted and that their value in promoting behavior change will be reduced [Fogg 2003, p.127]. Several of our users noticed contradicting feed items over time and it led some not to trust the system and the observations in their feeds. C3 told us that in the feed "you'll have three things there and maybe two of the three things contradict each other like in my mind. So it seems like because it's so vague it sounds like it all contradicts each other but maybe if it were a little more in the weekly correlation or something like on Mondays, this has been the case on every Monday for the last four weeks, that just makes more sense to me than randomly asserting that in general I'm eating more when I walk more." Having more specific details than the "vague" correlations was a common theme from our users. We had hoped that the graphs would have provided this additional data, but few users explored them in detail when encountering a feed item that they did not understand perhaps due to a general lack of graph literacy as noted by Galesic and Garcia-Retamero [2011].

S2 found that "information is repeating [in the feed] and I don't really understand how all the information is collected and how the correlations are computed." For some, this lack of understanding led to a lack of trust in the data. That observations often repeated on a daily basis discouraged frequent use because most of the same observations remained significant from day to day. Many participants reported wanting something "new."

Design Recommendation: Need for Reminders

Some participants suggested using the mobile phone notification system to alert users when new observations were available. C2 pointed out, "With the Android you get little notifications up top when you get something new like a new voicemail or text. Maybe something like that if there's a new correlation." C4 agreed: "The notification bar would work also. I think that would be good too. Something visual that would just be a reminder that you would see."

Other participants needed reminders for using the sensors. C4 said, "I've noticed that when I get text messages for reminders for doing things like I'll always do it. So if I was to get like a text message on a daily basis … in the morning saying hey, go on your scale or did you put your pedometer on, I think that would be extremely helpful." Our participants have clearly demonstrated the need for reminders and Fogg's early work using text messaging as triggers for health reminders [Fogg and Allen, 2009] should be applied even when there is a widget and application present with deeper interactions.

Our system required motivated users who were interested in digging deeper into data about their wellbeing. But our participants needed extra encouragement and including reminders could have dramatically changed the ways that they engaged with the system on a daily basis. This observation is in line with the Behavior Model from BJ Fogg where changing a behavior requires the motivation, the ability or knowledge to change, and triggers to remind the user about it [Fogg, 2003]. While our users were motivated to improve their wellbeing and knew how to do it (more activity, eating less, etc.), we were hoping that our observations would be this trigger. However,

they proved to be too passive especially when the items were on secondary home screens and users were lacking the explicit triggers they needed to engage with the service.

By adding reminders to improve logging frequency, we hoped to improve the data quality in the system so that contradicting observations would become much rarer in the full study.

5. FULL SYSTEM

Based on our experiences with the pilot study, we made several enhancements to the system for the larger study. The two primary weaknesses in the pilot were the infrequent use of manual logging and the contradictory nature of some of the observations over time as new data points were added. We hoped that by making changes to the system we could increase engagement, and thus the quality of the data and observations that were presented to users.

The first major change included adding reminders, in the form of status bar notifications on the phone. The first time that the Mashups application was opened users were presented with a list of the data that they could manually log (e.g. Food, Mood, Pain) and were able to set reminders that would appear each day at the same time. These times could be modified later by going into the application's settings screen. When the time of a reminder came due, a notification was silently placed into the status bar on the phone. We did not want the system to be interrupting, so no sound or vibration was used. However, the status light on the phone would blink to catch the user's attention on the next occasion that they had time to interact with the phone. We hoped that this would be enough to catch users' attention and encourage them to log each day when they might have otherwise forgotten. This was confirmed by the data as discussed in Section 7. We also included status bar notifications that appeared whenever a new observation was found for a user. These notifications led straight to the observations list when clicked.

The second change involved the addition of new data streams to the system to better reflect aspects of context and wellbeing that would be most important to users based on the pilot study and based on discussions with medical and insurance professionals. We added the ability to manually log Mood and Pain as well as added Weather data as an additional contextual sensor based on the current city-level location. This brought the total number of data streams to nine as shown in Table 1. The Fitbit and WiThings scale could automatically upload step count, weight, and sleep data. The phone automatically captured location, weather, and calendar free/busy data. And users could manually log Food, Mood, and Pain each day.

Logging mood consisted of four 7-point sliders corresponding to standard measures of mood: happy/sad, tired/awake, unwell/well, and tense/relaxed. Logging food consisted of three 7-point sliders of eating a little/a lot, healthy/unhealthy, and eating mostly at home/mostly out. These measures were created to be very quick to log in just a second or two, while still capturing a variety of the aspects of eating that could be influenced in various ways by one's context.

We also wanted to make it easier for users to access all observations about a particular sensor, so that they could better understand what was affecting each specific area of their life. Participants in the pilot often had specific goals for behavior change and we wanted to support someone easily understanding what affected their sleep or weight without having to wade through other observations that were unrelated. The updated application provided a list of each sensor, a sparkline for the recent changes in that sensor's value, and the date it was last logged with the corresponding value as shown in Figure 3. Clicking on a sensor item allowed the user to see all observations related to that sensor as well as a bigger graph of that sensor's recent values.

5.1 Study Method

Sixty diverse participants were recruited using a professional recruiting agency to use the system in their daily lives for 90 days. Participants represented a diverse array of ages, occupations, education levels, and general wellbeing. Twenty participants lived in Chicago or surrounding suburbs while 40 were from the Atlanta region. We wanted a broad set of participants in order to understand how awareness of one's wellbeing could be useful for a variety of needs and with varying technical literacy. Some participants lived in subsidized housing while others lived in large suburban homes. Some had very low education levels while others had advanced degrees in

scientific fields. Some were up to 150 pounds (68 kg) overweight while others were at ideal weights or even underweight. Importantly, for comparison with the pilot study, we used the same recruiting agency and screener for the full study.

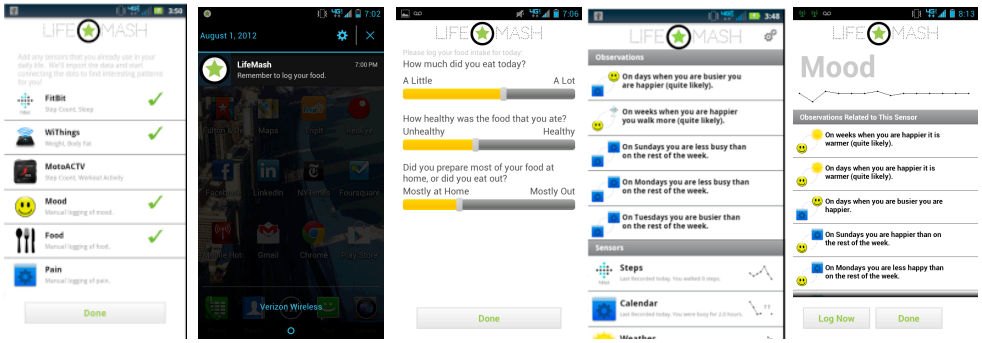

Figure 3: The mobile application for the main study. (2A) The user selects types of data to collect and sets the time for the reminder to appear in the status bar. (2B) Receiving a reminder notification for food logging, (2C) Clicking on the notification led to the simple food-logging screen in the middle. (2D) The observations could be reviewed at anytime but the system also added a reminder when a new observation has been found, below the observations is a list of each sensor with sparklines used to see recent changes. (2E) Clicking on a sensor will bring more details regarding this particular input, in this case the mood trend and related observations.

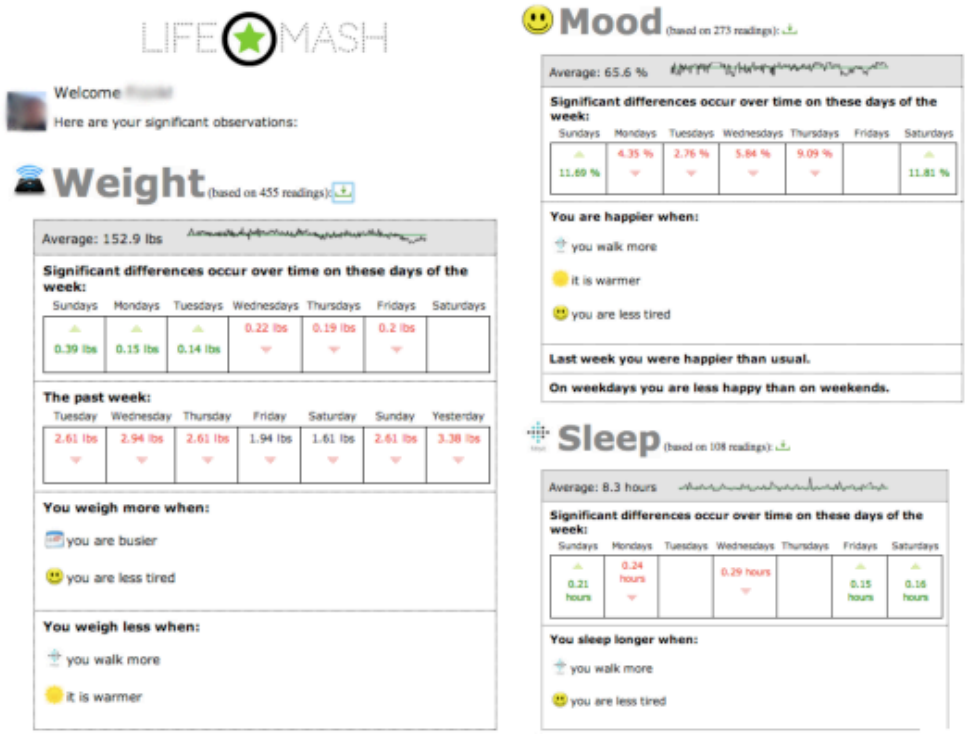

Figure 4: The web interface that was shown to users at the end of the study. It provided a clearer aggregation of the observations than the more constrained interface on the mobile device.

Researchers met with participants in their homes at the start of the study and set up the WiThings scale and Fitbit as well as installed the Mashups application onto the participants' primary phones. Participants were also given an initial questionnaire about their wellbeing and goals to complete. The final instructions to the participants (as in the pilot) included the request that they were to use this application like any other application they had downloaded from the market and that use was not required to participate in the study.

After 21 days, participants were sent a link to another online questionnaire. This covered open-ended questions relating to their use of the application, observations that they have found to be

useful or useless, and any behavior changes they might have made based on what they observed in the application. In addition, they completed the WHO-5 wellbeing [WHO, 2011] questionnaire that was also included in the initial survey.

After 90 days, a final questionnaire link was distributed with similar questions to the 21 day survey. Additionally, participants were sent a link to a web dashboard view of the observations that aggregated findings in an easier-to-read way (see Figure 4). A final semi-structured interview was conducted to elaborate on statements from the final questionnaire and to ask additional questions about the web dashboard.

Participants who completed the study (completed the 3 week and 3 month questionnaires and final phone interview) could keep the Fitbit and WiThings scale. Participants were also compensated with small gift checks to retail stores for submitting the questionnaires. No compensation was tied to the use of the system, and this was made clear to all participants.

In addition to the mostly qualitative data from the questionnaires and interviews, the mobile application was instrumented to collect usage data of each screen view in the application and this was reported to our server. Therefore, we were able to track when users were viewing their observations, manually logging, or viewing more complex features of the application such as the graphs or sensor-specific views. Quantitative analysis was performed to identify differences in use or wellbeing outcome based on age, gender, city, educational level, and other attributes. All qualitative data was combined into a large 1800+ note affinity to identify themes across participants. Both the quantitative and qualitative findings will be discussed in detail below.

6. FINDINGS

In contrast to the pilot, our participants demonstrated strong sustained engagement with the service. They consistently logged food and mood at much higher rates than in the pilot study, mainly due to the reminders that appeared on the phone. Due to this added data, the system was able to make more accurate and less contradictory observations and participants were able to both gain a better understanding about their wellbeing as well as make significant targeted behavior changes based on the observations that the system made.

6.1 Sustained Use

Our participants showed a much greater engagement with the system than was observed in our pilot study. This engagement covered all aspects of the system and did not show significant decline throughout the 90 days of the study. The dramatic increase in logging was the key to continuing engagement with the rest of the system, and the reminders likely contributed to this increase. Overall use can be seen in Figure 5 where each column represents a user and each dash represents an interaction with the application. It can be seen that most users that persisted beyond the first two weeks used the system quite regularly for the remainder of the 90 days. This is quite encouraging as other self-logging systems often see a sharp decline in use after 10-14 days [Burke et al 2008; Patrick et al, 2009].

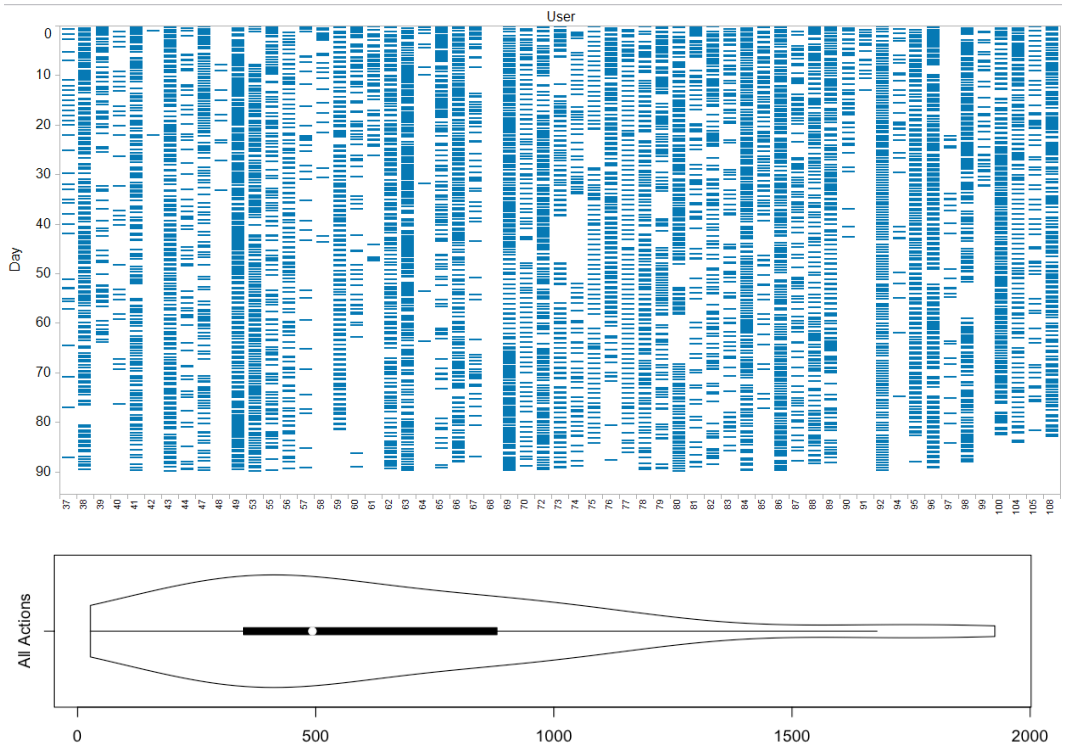

Figure 5: Use over time for each user. Each column represents a user and each dash is an interaction with the mobile application. Time follows from the top down for 90 days. The bottom graph shows the range in total number of screen views within the application across participants over the study period.

Use was consistent with regards to all demographics of our participants. There were no statistical differences in use between men and women, people in Atlanta and Chicago, by education level, or by age. This is quite encouraging as all ages and demographics were similarly engaged at very high levels, indicating that the presentation of these complex statistical relationships in natural language is a useful and understandable way to present wellbeing data to a broad range of people.

In our pilot study, the manual food logging was rarely used. In the first week, a few users tried it out, but after day seven, no more than two out of ten users logged food on the same day. After day 12, only one user sporadically logged food for the rest of the month as shown in Figure 6. This left us with an overall food logging rate of 12%.

This contrasts with the larger study with reminders enabled where 63% of users logged food each day in the first month. This percentage stayed consistent throughout the month, showing the power of simple reminders to promote sustained logging. Logging behavior was sustained beyond the first month as well with numbers between 50-70% each day during the second and third months of the trial. The average logging frequency for each data type can be seen in Figure 7.

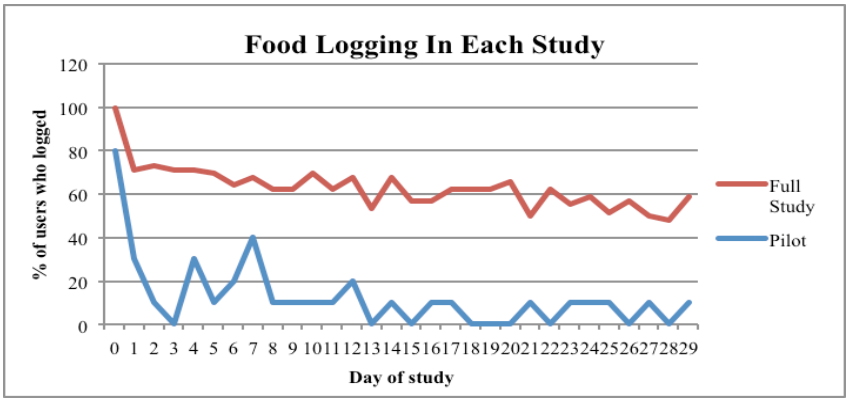

Figure 6: The rate of food logging behavior per day increased by more than 5x in the full study (with reminders).

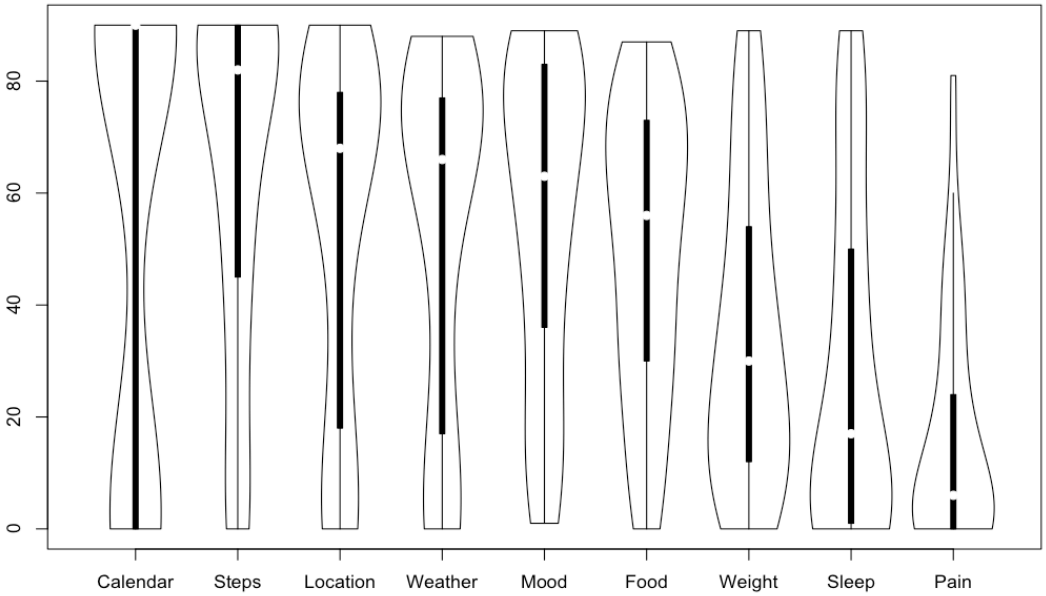

Figure 7: Frequency of logging each data type in the system over the 90 days of the trial. Sleep and pain were the least logged, due to the cumbersome nature of using the FitBit to log sleep as well as the fact that most participants did not have recurring pain. The automatic context streams of Calendar, Location, and Weather were quite bimodal with many participants not using these sensors at all, and others who kept them on daily. Overall, engagement is much higher than what was observed in the pilot system (Figure 2).

Our users reported liking the reminders as they recognized how easy it would be to forget to log. A28 stated that the best feature of the app was: "the reminder feature because half the time I forget to login my information." Also, A37 told us, "I like that it reminds me to add info in my notification bar. I would surely forget otherwise." C10 agrees: "The online reminders are awesome, makes it so much easier to keep track of stuff because I can get absentminded and lose track of what to do. The questions are easy to answer and make for a quick Q and A." This brings up the important point that the action required to enter data after clicking on the reminder should be as quick as possible. In our case, just a few 7-point scales, in comparison to other food-logging solutions that require complex logging of every food item eaten that can take ten minutes or more per day.

C20 spoke of the power of logging to increase reflection each day. "I like that it asks me to log my food and my mood daily. It makes me more conscious of how I'm feeling as well as what I'm eating throughout the day." The more often someone logs, the more opportunities for reflection they get.

The reminders also encouraged our users to interact with the observations that were created about their overall wellbeing by clicking on the notification to see newly significant observations. Viewing the full list of observations increased significantly from the pilot to the main study with this feature added.

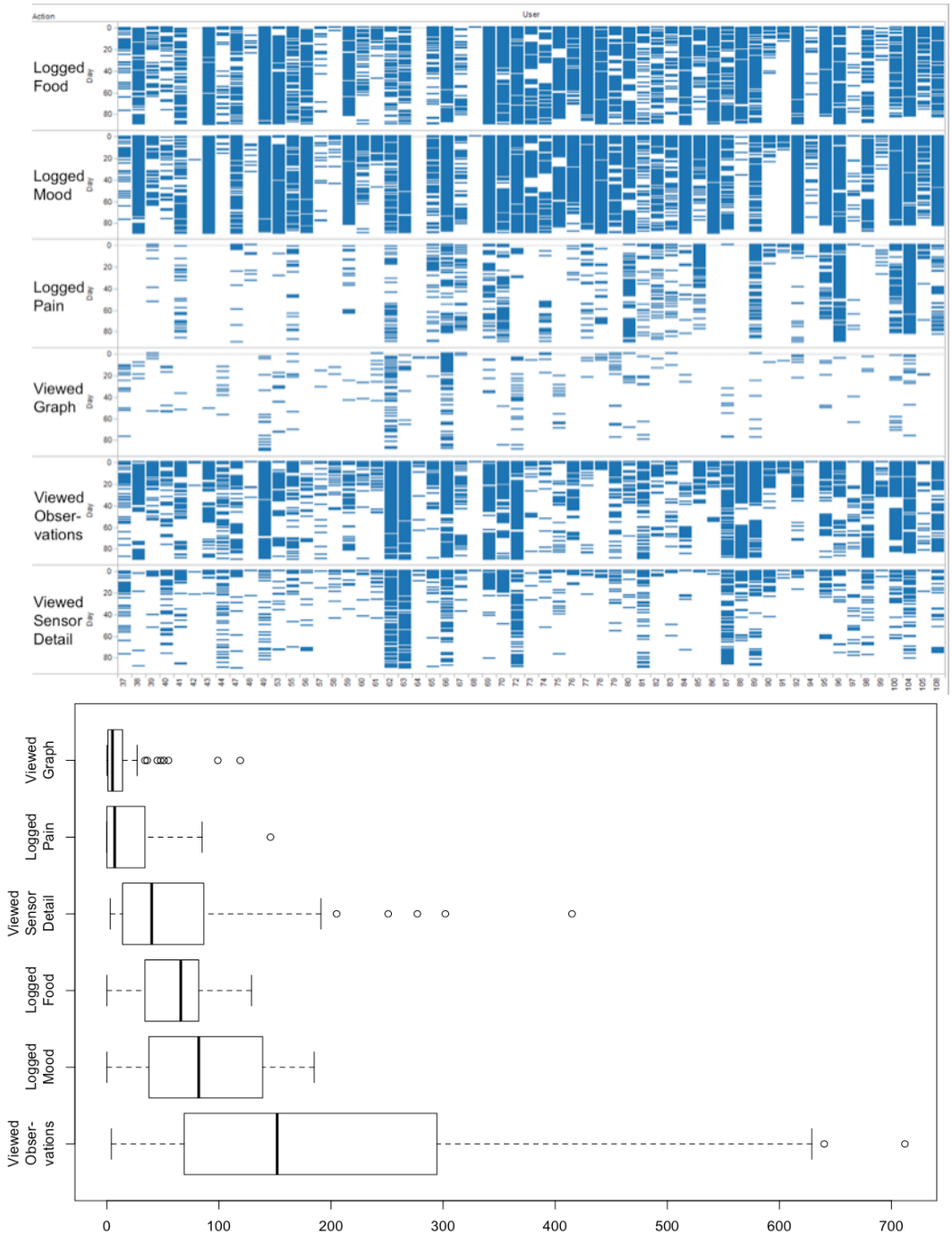

Figure 8: Usage of each feature of the mobile application over time, and overall. In the top figure each user is a column and time runs down from day 0 to day 90 in each section for a feature of the system. The bottom figure shows the raw count of our users' interactions with each feature during the 90 days of the study.

Many participants commented that the Mashups application became a part of their daily routine. User 4 from Chicago told us, "I learned that this is an app that is useful enough for me to use on a daily basis like Facebook. Most apps on my phone I don't use on a daily basis." User 16 from

Chicago had previously been in a health study from Northwestern where she was recording detailed food and health information into a smartphone. She described how time consuming this was and how quick it was to interact with the Mashups system each day. She noted that Health Mashups was much "easier to stick with." She also appreciated viewing the long term trends in the Mashups system and this kept her coming back week after week to learn more about herself.

Overall, the notifications in the system served to continually engage users and remind them to enter data into the system. This data then provided more accurate observations and the notifications about new observations brought users back to the system yet again. This increased interaction greatly improved both the amount and quality of data in the system as discussed below.

6.2 Quality of Observations

The increased logging not only allowed the users to reflect more on their own, but provided better data to the analysis engine and thus more accurate observations were presented to users, combating the problem of contradictory information seen in the pilot. The more days that have data points from multiple data streams for a given user, the more accurate correlations across sensor streams we can provide. For example, in the month of our pilot study, users averaged only a single day that had both Food and Weight logged, while in the full study participants averaged 9 such days in the first month and 21.3 over the full 90 days. For the Steps and Food combination, users averaged 2.6 days in the pilot study, 14.8 days in the first month of the full study, and 37.9 days in the full three months. This increased amount of data lead directly to the ability to perform a better statistical analysis and find significant correlations between the data streams. This was not something that could be reliably calculated between the manually logged sensors in the pilot study and the reminders were the driving force to encourage this additional logging.

	Food	**Mood**	**Pain**	**Steps**	**Weight**	**Sleep**	**Calendar**	**Location**
Mood	26/48/70							
Pain	0/6/21	0/6/22						
Steps	11/30/64	11/36/73	0/2/21					
Weight	4/18/37	3/18/39	0//1/9	6/18/43				
Sleep	1/7/34	1/9/33	0/0/6	1/16/49	0/5/17			
Calendar	0/24/51	0/22/56	0/0/14	0/22/73	0/6/23	0/1/24		
Location	2/30/58	1/34/65	0/0/13	0/36/62	0/13/32	0/4/27	0/33/64	
Weather	2/30/56	1/31/64	0/0/13	0/36/61	0/12/31	0/4/26	0/17/62	15/66/77

Table 2: 25th percentile, median, and 75th percentile of data points per user between sensors. These numbers represent the number of days when a user logged both the sensor from the row and column of each entry on the same day, thus creating a data point for the correlation.

We also were interested in exploring how unique the significant observations would be across participants. When we began this work, our hypothesis was that the ways in which context influences wellbeing would be highly individually different. We assumed that some people might walk farther on warm days while others might walk less. Or some people might have less pain when they walk more while others might have more. In order to see if this was the case, we analyzed all 450 statistically significant sensor-to-sensor correlations for the 60 participants in our study. These observations were quite stable over time with an average duration of 22 consecutive days where an observation was statistically significant. This is quite encouraging, as once an observation crosses the threshold of significance, it typically remains that way for some as new data is added. Even if a user tries to target that behavior for change (e.g. deciding to walk more on hotter days after the system informed them that they did not walk as much on these days), it should take several weeks for this to lose its statistical significance given the need for a good number of new data points to be added to weaken the strong correlation. This is supported by the 22-day average duration of the observations. Future work can explore ways to visually

demonstrate how behavior changes are weakening correlations over time or ways to develop goals to "remove" particular correlations that are negative to one's wellbeing.

In order to better understand the types of correlations that were most common across our user population, we looked at the correlation coefficients for each combination of sensors in each significant correlation that existed. The results can be seen in Figure 9. It is interesting to note that nearly all pairs of sensors had both positive and negative correlations for different participants. This confirmed our hypothesis that the interactions between wellbeing and context are specific to individuals. In fact, the only correlation that was universally positively correlated was the correlation between amount of food eaten and mood. The vast majority of differing correlation directions validates the premise of the system, that is, individuals need to capture their own wellbeing data and context to learn about the specific ways that their wellbeing is impacted by the activities in their lives. No simple set of universal guidelines exists.

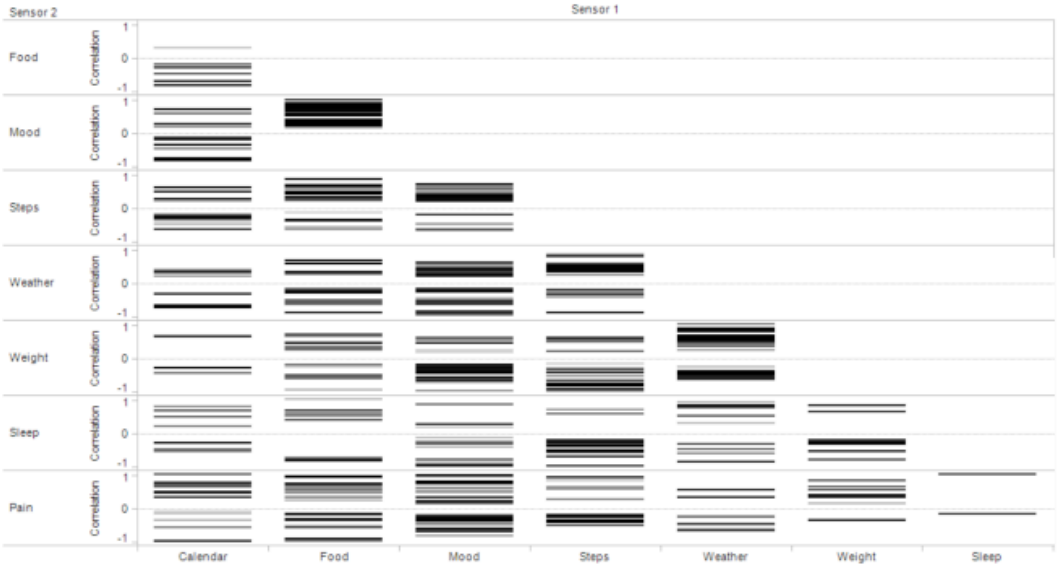

Figure 9: The multi-sensor correlations that were found to be significant after 90 days plotted against their correlation coefficients. Every pair of sensors (except for amount of food and mood) that produced significant results has both positive and negative correlations. Each band is a specific observation for a user in the study. The darkness of the band represents the p-value for the correlation (light = 0.05, dark < 0.01).

6.3 Increased Self-Understanding

Through viewing the observations that were displayed in the system, our participants were able to gain a deeper understanding of their wellbeing and the aspects that affected areas of their lives as diverse as mood, sleep, food intake, daily step count, and weight. Participants could see relationships that they had not expected and became a bit more introspective about their own wellbeing.

A8 broadened her view of particular health issues such as weight loss: "It made me realize that I need to stop looking at health issues as one entity." A26 learned "that everything is related. Mood, food, weather...all those elements are a factor in my well being."

Participants were able to learn their true activity and sleep patterns from the data on the sensors themselves and view trendlines that were displayed in the system over time. On top of this, they learned correlations based on actual data that represented measured fact instead of just their pre-existing hunches and speculation about how various aspects of wellbeing and context were related. For some, these observations served to confirm suspicions that they had, while others learned totally new insights about their wellbeing. A20 told us, "I do find [the observations] useful because it reminds me what my workout habits and eating habits really are like instead of me guessing. Takes a lot of guesswork out and actually makes me reflect on my day." A34 "didn't know that I was not as active as I thought I was. On the days when I didn't run or walk I realized that I didn't even cover a mile a day and was horrified!"

Specific observations helped to teach participants new patterns about their lives beyond just the sensor readings. C18 found that he consistently was less active on a certain day of the week. A37 learned something that was opposite to her expectations. She was surprised that for her, walking more in a day reduced her recurring pain. She thought it would be the other way around and previously was reducing her activity. “The info on when I walk more I'm in less pain really helped with my back. Made me realize I should exercise more and it dramatically helps with my pain levels.”

Some observations showed specific new connections between areas of participants’ lives that they did not know were related. A8 saw that she was happier on days when she ate more. She described that to us as “one of the most *mind blowing* things” because she “had never really associated both so closely or paid attention to them before.” Likewise, C13 was trying to lose weight and discovered that weight loss significantly impacted her mood. A2 learned more about the importance of sleep in her life. “I was able to see that on the days I slept less I was less happy and less motivated.” This made her realize that “I need sleep in order to function and live a healthier life.”

As a result of using the system, participants walked away with a lasting understanding of the factors that impact various aspects of their wellbeing. C16 now “consider[s] all components which affect my overall health much more.” C17 told us that “It's interesting to see how things correlate. And it makes you reflect on your day and be aware of what you've been doing and feeling.” Several other users spoke of the way that the system changed their focus to the long-term trends instead of the day-to-day fluctuations in specific sensors. A24 told us, “I love the observations. It helps me to understand myself from a birdseye view.” A42 talked about observations “making sense” to her and that “having quantifiable data to confirm was awesome.”

The increased self-understanding that people gained throughout the 90 days of the trial greatly improved their ability to see how diverse aspects of their wellbeing were related and the effects of context over time on their eating, weight loss, mood, or sleeping patterns. However, understanding is only the first step. Applying what was learned to make changes to patterns in daily life is the ultimate goal and many of our participants were able to make this leap.

6.4 Behavior Change

Beyond just increasing self-awareness, the observations allowed people to focus on specific behaviors to change in their daily lives. These types of change included focusing on a specific day of the week where their performance in some area was lacking, or focusing on a specific contextual aspect that impacted a particular area of their wellbeing.

The most basic area of focus occurred around specific days of the week. Oftentimes, participants would receive an observation that they walked less or ate more on a particular day of the week and in subsequent weeks they focused on being a bit more vigilant about this aspect of their wellbeing on that day of the week. A29 noticed that on Mondays he gets the least amount of exercise. That made him more conscious that he should get out and exercise more on the following Monday. C20 told us that “looking at the summary showed me that I am happier, but eating more on the weekends and therefore weighing more come Monday morning. This showed me that if I really want to lose weight, I need to be focusing on my eating habits on the weekends.” By focusing on eating less or walking more on specific days, making changes to behavior became specific and actionable.

Other focus areas were chosen based on correlations that appeared between two sensor streams. C2 noticed several observations related to activity and sleep. She saw that she slept better on days when she walked more and then started looking in her calendar to see what she was doing on the days when she did not walk as much. This prompted her to reschedule her activities so that she could be consistently more active and thus be able to sleep better. A30 was also concerned about sleep: “I learned that I should walk more and eat lighter meals to avoid sleeping longer.” A24 saw the observation that “I am happier when I walk more and I am less tired.” To her, “this was a great reminder that sacrificing sleep and exercise will not help anything I am trying to do or accomplish.” Because a specific correlated aspect of a user’s context or wellbeing was made

explicit in the observations in these cases, it was easy for participants to focus on these aspects to improve their wellbeing in the areas that mattered to them.

Another area that many participants focused on was increasing activity in general, and they could use the historical data presented in the system to monitor overall trends. Several participants spoke of walking more so that they could achieve their 10,000 steps per day goal or trying to get their weight below a certain level. The mix of showing the statistical observations as well as the current value and trendline for each of the data streams helped people to easily make these changes.

Overall the system provided a way to focus on specific behaviors in specific contexts that could benefit from change. Instead of just generally trying to "eat better," "walk more," or "lose weight," our participants could make specific focused changes to the aspects of their life that most impacted the behavior that they wanted to change. C19 summed this up: "I am a grad student who is overwhelmed and [the observations] helped to reflect on my life. They allowed me to take inventory and think about what I should change."

6.5 "Obvious" Observations

Overall, our participants were able to integrate the Health Mashups system into their lives and make many positive changes to improve their wellbeing. However, we also received many valuable comments on how some observations were not as useful as others. While most participants were able to apply the observations to create positive changes, others like C3 noted in the 3-week interview: "Does it really matter that I walk more on Tuesdays? Or that I eat more when I sleep less. Maybe as I use it more, it will make sense. But right now it isn't making a lot of sense for me." This raises a tension that some of our participants discussed between telling people what they already know and think is "obvious" versus providing observations that are educating and perhaps a bit more prescriptive. Some users mentioned wanting a rephrasing such as "Try to walk more on Tuesdays" that would give them specific goals instead of stating "the obvious" such as "You walk less on Tuesdays." However others appreciated that the system is not "judging" and does not "tell you what to do." Some participants noted that the system was not capturing all of the important context surrounding their behavior. C01 discussed the lack of support in adding additional information: "[The system] doesn't really take into account any outside factors. Like something [that the system doesn't capture] happens and that's why you're in a bad mood. … It's just defined rules and if it doesn't fall in that, then it doesn't comply." We will return to some of these issues in the discussion below.

6.6 Outcomes

At the end of the study our participants had an increased understanding about how their life context impacts their wellbeing and told us about many positive behavior changes that they initiated based on the information provided in the system. While we cannot quantify exactly how much of the behavior change was due to the presentation of observations over just using the scale and fitbit, the fact that many users experienced quite positive changes in wellbeing is promising.

Many of our participants had weight loss goals that they wanted to meet. Thirty-six of the 60 users in the study lost weight, at an average of 5.3 pounds (2.3 kg) over the course of the summer. Most other participants were already at a healthy weight and did not have a significant increase or decrease in weight throughout the study period. When asked about how they made these changes, participants reported focusing on specific, easy-to-target observations and making changes to these aspects of their lives as discussed in the Section 6.4.

We also observed a large and significant improvement in mood throughout the 90 days of the study. Forty of the 60 participants showed significant positive changes in mood, on average rising 29 percentage points on our sliding scale. From the qualitative data, this seems to be in some part due to understanding the correlations between food, sleep, activity and mood over time and making changes that contributed to being in a better mood. Improving mood through increased self-understanding is a rich area for further exploration, as we currently do not fully understand all that might have led to this significant increase over the 90 days. Seasonal changes in mood do

exist (e.g. [Kasper et al, 1989], [Howarth and Hoffman, 1984]), and future work would need to separate any seasonal change from changes due to the use of the system.

Along side these quantitative changes in weight and mood, we observed a significant increase in WHO-5 wellbeing scores [WHO, 2011] between the initial interview and the 3-week interview ($t(55) = 3.29$, $p < 0.003$). These scores remained higher from the 3-week interview until the end of the study showing sustained improvements in wellbeing from the beginning to the end of the 90-day study ($t(50) = 1.97$, $p < 0.05$). This demonstrates the potential for systems that encourage reflection on holistic impacts of context on various aspects of life to make a significant change to overall wellbeing, again with future work utilizing control groups needed to understand the effect that the generation of observations contributed to this increase.

7. DISCUSSION

We have shown how our users were able to build an awareness of certain contextual wellbeing patterns in their lives and focus their change efforts based on this new awareness. We have also shown how these effects are very personally unique, with different participants having a different set of significant observations that were often the opposite of those from other participants. This type of system can help people to understand their own lives, similar to Li et al's [2011] discussion of the "discovery" phase. As Li states "personal informatics tools are not designed with a sufficient understanding of users' self-reflection needs," and we have shown how paying attention to self-reflection as a main design goal can lead to a successful solution that allows users to learn deep insights about how their context impacts a variety of aspects of their wellbeing over the long term.

In their book "Nudge," Thaler and Sunstein [2008] argue that with small modifications we can influence choices about health and wellbeing in ways that will make us better off. They describe two different systems, the "Reflective System" and the "Automatic System," that help us making everyday choices. In particular, the reflective system might be influenced with an approach like ours. By helping users to notice and reflect on observations about their lives, we have seen how they can be "nudged" to make small, targeted changes that can quantitatively improve their wellbeing over time (e.g. by lower weight, increased mood, or increased WHO-5 wellbeing scores). The metaphor of a "nudge" is rather appropriate, as we are not telling people what to do or stating what is good or bad, rather we simply point out what is statistically significant about their lives and leave it to them to decide to make a change.

However, there is still an unsolved tension between telling people what they already know and think is "obvious" versus providing observations that are educating and reinforcing. Many of our participants saw some of the observations as obvious, such as being happier on weekends or gaining weight on Saturdays. For them, this data simply made sense with their lives, whether they had previously considered it or not. Discovering how to best present this data in a way that makes clear and actionable observations prominent is still an open question. While most of our participants eventually got beyond the "obviousness" of some observations and focused on ones that they had not anticipated, in many cases it was the obvious observation that was previously unstated that was the most fertile area for change in one's life. Future systems in this domain should take this into account, perhaps by allowing users to permanently hide observations that are too obvious, are uninteresting, or unrelated to current goals or by adding additional information that users might not find to be so obvious. Here, machine-learning approaches that have been used in recommendation systems (e.g. [Adomavicius and Tuzhilin 2005]) might be useful to more efficiently filter and further personalize the observations over time.

It is clear that a wide variety of contextual variables affect wellbeing and that capturing as many as possible will give users the best and most relevant insights into their own lives. We are still missing some aspects that users wanted to log such as water/caffeine/alcohol intake or other activities such as cycling or swimming. As Consolvo et al state in their design guidelines [2006], it is important to be able to cover as many types of activity as possible. This is especially true when performing statistical analysis on the data as missing activities such as swimming (when relying on pedometer data for activity) will provide quite different views on daily activity level and could skew correlations or day of week deviations if many of these types of workouts occur. Even with

the lack of additional sensors or manual activity entry, we were still able to make statistically meaningful observations that made sense to our users and related to their daily lives. Thus, having every possible input is not critical. But every user comes to the system looking for something different with unique wellbeing goals, whether it is improving sleep quality, reducing pain, or trying to be more active Each of these areas requires different types of sensing or logging to accurately capture the appropriate activities or contextual variables that can affect it.

Overall, our participants were quite happy with the system and the observations were able to provide new insights into their wellbeing that they were not able to make on their own or with existing wellbeing monitoring tools such as the Fitbit or WiThings websites, as shown in the pilot study. These findings from the pilot led to us not using an explicit control group for the larger study, as we saw the lack of depth in the types of insights that the existing systems can give. We see the importance of bringing Quantified Self-style logging and analysis to those without the statistical and analytical skills of many in the QS community and found that by providing observations as natural language observations, all users were able to understand how various aspects of their lives were interrelated, without the need to understand and interpret complex graphs or understand statistical terminology.

8. CONCLUSION

We have shown how a system can aggregate multiple aspects of wellbeing data and context for an individual over time and that the interactions between these are highly individually different. By surfacing significant observations between the data streams, our participants were able to become more aware of what impacted their mood, food intake, sleep patterns, and activity as well as create targeted behavior change strategies based directly on the significant observations shown to them by the system.

Our pilot system showed that reminders and other notifications were necessary to increase engagement to a point where enough data would be provided such that observations would be statistically significant and remain significant over time. The larger study showed how our participants were able to use the observation data in their daily lives to focus their change efforts and how the reminders provided the increased engagement that was necessary for a variety of meaningful observations to be found. Most importantly, once enough data was provided to the system, the observations that were generated for our participants were easily understandable, even by users with a wide range of education and technical exposure. They were able to understand complex relationships in their wellbeing and perceive that their focused change efforts were successful based on these observations to improve their overall wellbeing. Achieving high levels of engagement is critical for systems such as this one to gather enough data to make meaningful observations and to encourage reflection on the statements that the system makes. Increased engagement is especially important, as we have shown that these interactions between wellbeing and context are highly variable between individuals, necessitating a large amount of data to be calculated for each user.

We see the Health Mashups system as a first step towards exploring the rich personal connections between wellbeing and context over time. We expect this work to open the door to future work that explores other contextual attributes as well as other wellbeing measures. There is a large amount of future work in exploring the most relevant sensors and contextual streams that will give people the most useful information about what is affecting the aspects of wellbeing that they care about the most. And there is additional work in more fully exploring changes to mood or other wellbeing measures that systems like this can enable.

There is also room for research in the presentation of the significant observations. We chose natural language as we wanted an output that could be understood by people across education levels and graph literacy. This seems to have worked fairly well, but sometimes provided results that participants found to be "obvious." Learning which observations to show and how to convey not only the attributes involved but also the strength of the correlation or even trying to infer order in the presentation (e.g. the night's sleep came before the "happy" day, so we know that any causality cannot be the opposite direction) could be interesting research topics. With more sensors and contextual attributes comes longer lists of observations. Discovering ways to properly sort or

filter these lists based on the user's interest will become critical to the wide deployment of these types of services. Supporting explicit goal-setting/tracking or gamification around removing "negative" observations over time are also possible future directions for this line of work.

Systems such as Health Mashups show the power to increase awareness of overall wellbeing and highlight specific and actionable areas for targeted behavior change. We have shown how individually unique interactions between wellbeing and context can be automatically discovered and how, with enough data provided, these observations can be easily understood and used to create positive, focused changes to improve wellbeing. We hope that the findings and design inspiration from this work can open research into new ways to build awareness about the contextual effects of wellbeing over time and to encourage specific focused changes to improve wellbeing.

ACKNOWLEDGMENTS

The authors would like to thank Wireless@KTH for funding the pilot study and Humana for funding the full study. We would also like to thank all of our participants in both studies for their time and valuable feedback to help us to continually improve the concept and understand how this type of analysis can produce positive changes in wellbeing. We would also like to thank Cristobal Viedma for his work on the implementation of the pilot system.

REFERENCES

ADOMAVICIUS, G. AND TUZHILIN A. Toward The Next Generation Of Recommender Systems: A Survey Of The State-Of-The-Art And Possible Extensions, IEEE Transactions On Knowledge And Data Engineering, 17, NO. 6 (2005): 734-749.

ALBERS, D. J., HRIPCSAK, G., "A STATISTICAL DYNAMICS APPROACH TO THE STUDY OF HUMAN HEATH DATA: RESOLVING POPULATION SCALE DIURNAL VARIATION IN LABORATORY DATA" PHYS. LETT. A 374 (2010) 1159-1164

ANCKER, J.S., AND KAUFMAN, D. RETHINKING HEALTH NUMERACY: A MULTIDISCIPLINARY LITERATURE REVIEW. J AM MED INFORM ASSOC. 2007;14:713–721. DOI 10.1197/JAMIA.M2464

ANDERSON, I., MAITLAND, J., SHERWOOD, S., BARKHUUS, L., CHALMERS, M., HALL, M., BROWN, B., AND MULLER, H. 2007. Shakra: tracking and sharing daily activity levels with unaugmented mobile phones. Mob. Netw. Appl. 12, 2-3 (Mar. 2007), 185-199.

BECH P, GUDEX C, STAEHR JOHANSEN K. The WHO (Ten) Well-Being Index: Validation in Diabetes. Psychother Psychosom 1996; 65: 183-190

BUPA 2011, http://www.bupa.co.uk/individuals/health-information/directory/o/child-obesity

BURKE, L., SEREIKA, S., MUSIC, E., WARZISKI, M., STYN, M., Stone. A. Using Instrumented Paper Diaries to Document Self-Monitoring Patterns in Weight-Loss. Contemp Clin Trials. 2008 March: 29(2): 182-193.

CDC 2010, U.S. Obesity Trends. Centers for Disease Control and Prevention. http://www.cdc.gov/obesity/data/trends.html Accessed 09/06/11.

CHRISTAKIS, N. AND FOWLER, J. The Spread of Obesity in a Large Social Network Over 32 Years. The New England Journal of Medicine. V 357:370-379, No 4. July, 2007.

CONNELLY, K., FABER, A., ROGERS, Y., SIEK, K., AND TOSCOS, T. Mobile applications that empower people to monitor their personal health. Elektrotechnik und Informationstechnik 123(4): 124-128 (2006)

CONSOLVO, S., KLASNJA, P., MCDONALD, D., AVRAHAMI, D., FROEHLICH, J., LEGRAND, L., LIBBY, R., MOSHER, K., AND LANDAY, J. 2008. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In Proc. UbiComp '08.

CONSOLVO, S., MCDONALD, D., AND LANDAY, J. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In Proc. CHI '09.

CONSOLVO, S., EVERITT, K., SMITH, I., AND LANDAY, J. 2006. Design requirements for technologies that encourage physical activity. In Proc. CHI '06.

CONSOLVO, S., MCDONALD, D., TOSCOS, T., CHEN, M., FROEHLICH, J., HARRISON, B., KLASNJA, P., LAMARCA, A., LEGRAND, L., LIBBY, R., SMITH, I., AND LANDAY, J. 2008. Activity sensing in the wild: a field trial of ubifit garden. In Proc. CHI '08.

EMMONS, R. A., & MCCULLOUGH, M. E. (2003). Counting blessings versus burdens: Experimental studies of gratitude and subjective well-being in daily life. Journal of Personality and Social Psychology, 84, 377-389.

FOGG, B.J. AND ALLEN, E. 2009. 10 uses of texting to improve health. In Proc. Persuasive '09.

FOGG, B.J. Persuasive Technology: Using Computers to Change What We Think and Do, San Francisco, CA, USA: Morgan Kaufmann Publishers, (2003).

FOGG, B.J.. 2002. Persuasive technology: using computers to change what we think and do. Ubiquity December, 2002..

GALESIC, M., & GARCIA-RETAMERO, R. (2011). GRAPH LITERACY: A CROSSCULTURAL COMPARISON. MEDICAL DECISION MAKING, 31, 444–457.

GEIGERENZER, G.,GAISSMAIER, W., KURZ-MILCKE, E., SCHWARTZ, L., WOLOSHIN, S. 2007. HELPING DOCTORS AND PATIENTS MAKE SENSE OF HEALTH STATISTICS. PSYCHOLOGICAL SCIENCE IN THE PUBLIC INTEREST NOVEMBER 2007 VOL. 8 NO. 2 53-96

HOWARTH, E. AND HOFFMAN, M.S. (1984) "A MULTIDIMENSIONAL APPROACH TO THE RELATIONSHIP BETWEEN MOOD AND WEATHER," BRITISH JOURNAL OF PSYCHOLOGY , NO. 75, PP. 15-23, 1984.

KAHN, E.B. et al. The effectiveness of interventions to increase physical activity. A systematic review. American Journal of Preventive Medicine 22, 73-107 (2002).

KASPER S, WEHR TA, BARTKO JJ, ET AL. (1989) EPIDEMIOLOGICAL FINDINGS OF SEASONAL CHANGES IN MOOD AND BEHAVIOR. ARCH GEN PSYCHIATRY 46: 823±833

KLASNJA, P., CONSOLVO, S., AND PRATT, W. 2011. How to evaluate technologies for health behavior change in HCI research. In Proc. CHI '11.

LI, I., DEY,A. AND FORLIZZI, J. (2011) Understanding My Data Myself: Supporting Self-Reflection with Ubicomp Technologies. In Proc UbiComp 2011.

LI, I. Personal Informatics and Context: Using Context to Reveal Factors that Affect Behavior. PhD Thesis. 2011.

MAITLAND, J., SHERWOOD, S., BARKHUUS, L., ANDERSON, I., HALL, M., BROWN, B., CHALMERS, M., & MULLER, H., "Increasing the Awareness of Daily Activity Levels with Pervasive Computing," In Proc. Pervasive Health '06.

LIPKUS, I. 2007. Numeric, Verbal, and Visual Formats of Conveying Health Risks: Suggested Best Practices and Future Recommendations. Med Decis Making September/October 2007 vol. 27 no. 5 696-713.

MAMYKINA, L., MYNATT, E., DAVIDSON, P., and GREENBLATT, D. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). ACM, New York, NY, USA, 477-486.

Medynskiy, Y., Mynatt, E. 2010. Salud!: An open infrastructure for developing and deploying health self-management applications. In Proceedings of Pervasive Health 2010: 1-8

PATRICK, K., RAAB, F., ADAMS, M., DILLON, L., ZABINSKY, M., ROCK, C., GRISWOLD, W., NORMAN, G. A Text Message-Based Intervention for Weight Loss: Randomized Controlled Trial. J Med Internet Res. 2009 Jan-Mar; 11(1):e1.

PROCHASKA JO, VELICER WF. The transtheoretical model of health behavior change. Am J Health Promot 1997 Sep-Oct;12(1):38-48.

RACHLIN, H. The Science of Self Control. Cambridge, MA, USA: Harvard University Press, (2004).

THALER, R. and SUNSTEIN, C. 2008. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

TOLLMAR, K., BENTLEY, F., Viedma, C. Mobile Health Mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. In Proceedings of Pervasive Health 2012.

van Eck, M., Berkhof, H., Nicolson, N., Sulon, J. 1996. The Effects of Perceived Stress, Traits, Mood States, and Stressful Daily Events on Salivary Cortisol. Psychosomatic Medicine 58:447-458 (1996)

WHO. WHO-Five Well-being Index (WHO-5), http://www.who-5.org/ 2011. Web. 10 Sep 2011.

"well-being." Merriam-Webster.com. Merriam-Webster, 2011. Web. 10 Sep 2011.

Received December 2012

Statement of Previous Research

We have previously published on the pilot version of the Health Mashups system in a Pervasive Health paper from 2012 (referenced in the submission). This paper was a technical exploration of the pilot system and discussed the aggregation of data from multiple sources and the specific statistical analysis that was performed on the data streams for the pilot study. This paper did not include the detailed qualitative findings from the pilot or any content on the complete system or the full study.

We also have a Note currently in consideration for CHI 2013 about the increased logging that was seen by adding reminders to the system between the pilot and full study. This note only focuses on this one particular finding and does not include any other data about the full system or other usage data from the study.

The main contributions of the current submission, not covered by previous publications include:

1) Detailed qualitative findings about what participants were able to learn about themselves from the system (both for the pilot and full studies) and the specific behavior changes this enabled.
2) Detailed quantitative findings about the use of the system in both the pilot and full studies.
3) The analysis of both positively and negatively correlated observations for most combinations of sensors and contexts across participants (Figure 9) thus confirming the need for a system that performs individual analysis.
4) The narrative of the entire project from pilot study, to design improvements, to full study and the increased engagement we were able to create by following this design path.