Mobile Sensing at the Service of Mental Well-being: a Large-scale Longitudinal Study

Sandra Servia-Rodríguez
Computer Laboratory
University of Cambridge, UK
sandra.servia-rodriguez@cl.cam.ac.uk

Kiran K. Rachuri
Samsung Research America, USA
k.rachuri@samsung.com

Cecilia Mascolo
Computer Laboratory
University of Cambridge, UK
cecilia.mascolo@cl.cam.ac.uk

Peter J. Rentfrow
Department of Psychology
University of Cambridge, UK
pjr39@cam.ac.uk

Neal Lathia
Computer Laboratory
University of Cambridge, UK
neal.lathia@cl.cam.ac.uk

Gillian M. Sandstrom
Department of Psychology
University of Essex, UK
gillian.sandstrom@essex.ac.uk

ABSTRACT

Measuring mental well-being with mobile sensing has been an increasingly active research topic. Pervasiveness of smartphones combined with the convenience of mobile app distribution platforms (e.g., Google Play) provide a tremendous opportunity to reach out to millions of users. However, the studies at the confluence of mental health and mobile sensing have been longitudinally limited, controlled, or confined to a small number of participants. In this paper we report on what we believe is the largest longitudinal in-the-wild study of mood through smartphones. We describe an Android app to collect participants' self-reported moods and system triggered experience sampling data while passively measuring their physical activity, sociability, and mobility via their device's sensors. We report the results of a large-scale analysis of the data collected for about three years from $\sim 18,000$ users.

The paper makes three primary contributions. First, we show how we used physical and software sensors in smartphones to automatically and accurately identify routines. Then, we demonstrate the strong correlation between these routines and users' personality, well-being perception, and other psychological variables. Finally, we explore predictability of users' mood using their passive sensing data. Our findings show that, especially for weekends, mobile sensing can be used to predict users' mood with an accuracy of about 70%. These results have the potential to impact the design of future mobile apps for mood/behavior tracking and interventions.

1. INTRODUCTION

Mental health illnesses are one of the most prevalent diseases worldwide. According to the National Institute of Mental Health (NIMH), over 40 million adults in the United States had a diagnosable mental illness in 2014 [5]. Mood disorders, which impact a person's emotional state, affect 20% of American adults at some point during their lifetime. Further, in any given year, 1 in 4 people in England and 1 in 5 people in Australia experience a mental illness [4, 1]. Although mental health problems are widespread, a substantial percentage of people do not receive treatment. Indeed, the World Health Organization (WHO) estimated that 35-50% of affected people in high income countries and 76-85% of affected people in low to medium income countries do not receive treatment [3].

Psychologists and social scientists have been studying mental health and well-being for decades [40, 16]. Traditional psychology methods to track mental well-being include pen and paper surveys, and controlled laboratory observations using a limited number of participants. However, the large penetration of social media sites and smartphones, together with recent advances on mobile sensing technologies, have allowed researchers to conduct studies using unobtrusive methodologies, reaching many more participants [35, 10, 45]. Although social media technologies have been demonstrated to be useful for tracking users' happiness, they usually show a partial view of peoples' lives: their *online* life. Sensor-rich mobile devices, such as smartphones/watches, on the other hand, have been proven effective in monitoring real-world activities and behaviors [14, 43]. However, most existing studies tracking mental well-being using mobile sensing have been conducted using a limited number of participants – indeed, a substantial percentage of existing work used less than 100 participants [36, 34, 30].

In order to overcome the limitations of low number of participants and unavailability of the complete view of users' daily lives, we developed a Android-based application for mood monitoring that blends mobile sensing with mood reports from users. By collecting data from the most commonly used physical sensors (microphone, accelerometer, location) and software sensors (text messages, phone calls) in smartphones at different duty-cycling rates, and by asking users to report their mood twice per day, our application has been able to obtain and provide users with useful insights about their moods; for example depicting the distribution of their mood reports over time or correlating it with sensor readings such as microphone ambiance detection.

The app was released to the public as a free download on *Google Play* in February 2013. We then conducted a campaign to attract participants that included a press release from one of the participating universities and the subsequent media coverage. After about three years on the *Google Play* store with over $40,000$ downloads and $18,000$ users, the app yielded an unprecedented amount of sensed and self reported data. *This is, to our knowledge, the largest dataset of mood and passive sensing data both in terms of partic-*

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052618

ipants numbers as well as duration. Our mobile application has been designed by an interdisciplinary team of computer scientists and social psychologists. Using the sensing capabilities of smartphones, and taking into account battery constraints and data transfer costs of users' devices, our aim was to identify sensor data features, and sensing/reporting rates that would enable the community to build efficient and effective psychological care applications. In this paper, we present the design and deployment of our smartphone application and a large-scale analysis of the collected dataset. In the data analysis we explore: (i) the link between inferred routines from smartphone sensor data and user's demographics, personality, well-being perception, and other psychological factors and (ii) if a diverse set of sensor signals can be used to track user's mood and happiness. Compared to existing studies, our work is the first to study the relation between passive mobile sensing, inferred routines, personality, and well-being using a large-scale data spanning thousands of users.

The main contributions of this paper are the following:

1. We report about our mood monitoring app and its use of mood reports and passive sensors to track user mood and report feedback for over its three years deployment.
2. We demonstrate how physical and software sensors in a smartphone can be employed to automatically and accurately identify routines of users. While various studies have reported similar findings, our results are at an unprecedented scale.
3. We show that these inferred routines are not independent from users' personality, well-being perception and other psychological variables, but in many cases, they exhibit a correlation with these factors: for instance, demographics such as gender, age, and education show a strong correlation with user's inferred routines.
4. We explore predictability of users' mood by using passive sensing data. We found that, especially for weekends, mobile sensing can be used to classify users' positive and negative mood with an accuracy of about 70%. This is likely to have positive implications on the design of future mobile phone applications for mood monitoring: we demonstrate that passive sensing data has the power of quite accurately approximate user mood reports thus making them unnecessary.

We believe the scale of the deployment and dataset both in terms of duration and number of participants make these findings robust to the noise introduced by "uncontrolled" manner in which users were recruited and the data collected. Indeed, our study is $10x$ bigger than similar studies and still maintains high accuracy.

2. OUR APPLICATION FOR MOOD MONITORING

In this section we describe our mobile application and the collected dataset. Our app has been designed by a team of psychologists and computer scientists to study subjective well-being and behavior. It collects self-report data through surveys presented on the phone via experience sampling. By default, in order to not burden the user [40, 15], the app sends only two notifications per day at random moments between 8AM and 10PM, at least 120 minutes apart from one another. Clicking on a notification launches a momentary assessment, which includes measures of current affect, and measures of individual aspects of behavior or context (e.g., physical activity, location, and social interactions). In addition to the notification driven surveys, the app also collects self-initiated surveys. These included longer measures of affect, and measures assessing multiple aspects of behavior and context. The app uses open sourced software libraries presented in [32] to periodically collect

(a) Affect grid. (b) Sample mood-related question. (c) Mood vs. microphone (noise level).

Figure 1: Screenshots of the App.

behavioral data from physical and software sensors in the phone (accelerometer, microphone, location, text messages, phone calls). The data collected through the app is stored on the device's file system and then uploaded to a server. To reduce data upload costs, the data is first compressed and then uploaded to the server only when the phone is connected to a Wi-Fi hotspot.

The app was designed to be a tool to facilitate self insight, providing feedback on how participants' mood relates to context and activity. In an effort to maintain user engagement over a period of weeks, participants could receive additional feedback by "unlocking" stages, each of which had a particular theme (e.g., physical activity, location, social interactions) that determined which behavior and context questions were asked in the self-reported surveys. Figure 1 shows some screens of the app. The screenshot in Figure 1a represents the affect grid [47] used to collect graphically the mood of the user: participants are supposed to tap on the grid to report their mood. The x-axis indicates the feeling in terms of its positiveness or negativeness while the y-axis indicates its intensity. Users answer a presented survey question by selecting a Likert rating as shown in Figure 1b. The screen in Figure 1c includes the feedback given to the user in terms of microphone intensity and mood in the last seven days, after they "unlock" a certain level. Different colors indicate the intensity. The app also asks profile related questions. The ground truth information was collected by asking questions, for example, on user's recent physical activity, location, after the user completes a survey. We present more details on the survey questions in Section 4.

2.1 Dataset

Our dataset contains users' sensed and self-reported data from February 2013, when the application was launched, until January 2016. Table 2 contains the description of the data collected that was used in our analysis. Over $40,000$ people downloaded the application after a successful media campaign. The demographics of the participants are reported later in the paper when we present analysis results. More than $18,000$ users used the app and provided data used for this study, although there are differences between the number of users for whom we have sensed data and self reports. For each sensor, we have data from at least $11,000$ users (Table 2). Figure 3 shows the complementary cumulative distribution (CCDF) of days for which we have collected at least one sensor data sample or the user filled at least one survey. For instance, for sensors reporting location, we have over $6,000$ users with at least three weeks of

		# users	# samples	avg (# days)
sensors	location	15,017	6,108,477	39.36
	accelerometer	18,657	16,770,659	22.22
	microphone	18,743	18,189,917	22.51
	text messages	11,589	10,750,608	16.82
	phone calls	11,470	9,154,372	23.25
self-reports	location	13,285	212,987	9.37
profile related surveys	demographics	11,181	11,181	NA
	personality	3,354	3,354	NA
	gratitude	3,070	3,070	NA
	health	2,941	2,941	NA
	sociability	2,700	2,700	NA
	job satisfaction	2,424	2,424	NA
	life aspirations	2,485	2,485	NA
	connectedness	2,420	2,420	NA
	swls	13,883	13,883	NA
experience sampling	affect grid	16,370	862,927	26.01
	PANAS	15,798	1,582,444	25.67

Figure 2: Data used in the analysis.

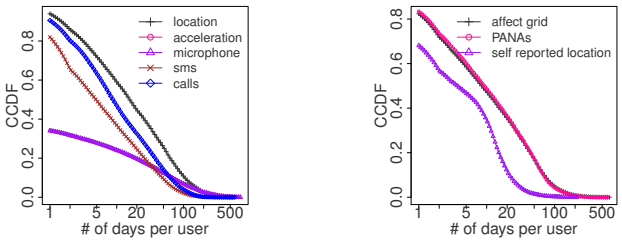


Figure 3: CCDF of user days for sensor & self-report data

data, and over $1,500$ users with over three months of data. Given the varied amount of data for each of the sensors and self reports, we present more details on the number of participants and days of sensor data used for each analysis in their respective subsections.

3. INFERRING ROUTINES FROM SENSOR DATA

In this section we study the strength of passively collected sensor data in understanding users' routines. We focus on routines that have an impact on user behavior, such as location type or activity level [34, 10, 45]. We consider five different sensors that indicate: the user's environment (location and microphone), activity (accelerometer), and sociability (messages and calls). Although routine detection has been explored by earlier work [27, 41], the scale and in-the-wild model of our study adds considerable noise.

The battery and network costs of mobile phone devices combined with sporadic turning off of the phone/app by users, made necessary to apply an extensive process of data cleaning. For each sensor, we identify days when data was sensed at least once an hour for a participant: we refer to these days as *high sensor coverage days*. Data from all high sensor coverage days constitute the final dataset used in this analysis. We now describe the participants lives as seen through the lens of the passive sensors that the app used.

3.1 Environment

Existing work reported that mood is often affected by environmental conditions [34, 45], so we first tried to determine the environmental context of a user using location and microphone sensors.

3.1.1 Location

Our application senses participants' location using GPS, Cell-Id and Wi-Fi at the rate of, an average, one sample per hour. As a matter of battery-efficiency, most of the measures were collected using Wi-Fi and Cell tower information. Although this provides a coarse location by trading-off accuracy for conserving energy, it is sufficient to determine the location context of the user, which we will show further in this section. Moreover, we pause sensor data collection when the battery level falls below 20% in order to not drain the battery rapidly when its already low. Pre-processing the full dataset to extract high sensor coverage days resulted in $4,590$ days of data from 700 participants ($3,297$ week days and $1,293$ weekends). Further, given the coarse-grained nature of the location data, like in earlier work [48], we assume that two locations are identical if their distance is less than $1,600$ meters, which may not necessarily be the case in highly dense urban environments.

The app asks participants to report their location context twice per day, choosing from: *Home*, *Work*, *Restaurant/Cafe/Bar*, *In Transit*, *Family's/Friends' house* or *Other*. The dataset includes $212,987$ self-reported locations from $13,285$ users. Since the sensed location is finer than the self-reported, we matched the sensed location with that of user reported to have a fine grained representation of user's locations through the day. We consider that there is a match between these if the location was sensed in the 15 minutes prior or subsequent to the corresponding self report. The data for this analysis is the aggregation of the matched locations for the 700 users with high location sensor coverage.

Number of locations visited per day: Given the coarse granularity of the sensed data, i.e., one sensed-location per hour, we consider a location *visited* if a participant spent at least an hour in that location, thus, it eliminates transient locations. Figure 4a shows the total number of locations, and unique locations, visited per day during week days and during weekends. We see that, on average, our participants visit more locations during week days (3.13) than on weekends (2.72), and visit 2.51, 2.12 unique locations on week days, weekends, respectively. On any given day, users are unlikely to revisit the same place twice unless this place is their home.

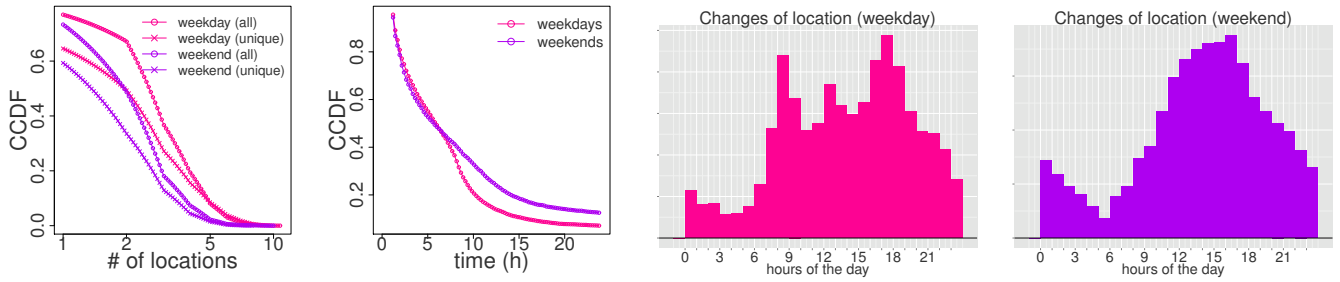
Time spent at each location & Location changes vs. time of day: Figure 4b shows that, on average, users spend more time at a place during weekends than during week days. This observation is coherent with the fact that they visit less locations during weekends. This means that users tend to be more "location active" on week days than on the weekends, i.e., change more locations. Figures 4c, 4d show that changes of location usually happen during day time. During week days, they are more likely around commuting hours (8am - 5pm) and at lunch time (12pm). On weekends, during day time especially from afternoon to evening.

Our findings are in-line with existing studies [27, 41], which reported that people spend most of their time at a few key locations, and location change peaks are different between week days and weekends. Although location was sensed at longer intervals using low-power sensors (Wi-Fi, Cell Ids) for saving energy, it still can be useful to understand the users' routines at a coarse-level.

3.1.2 Microphone

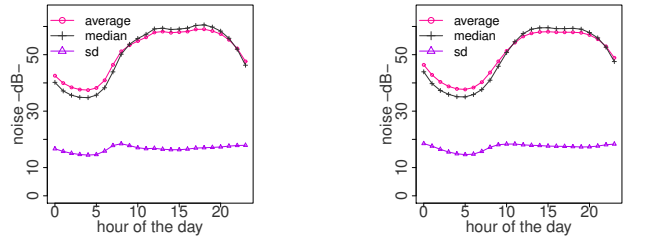
Our app captures noise level in the user's environment using the phone's microphone. To preserve privacy we do not save any raw audio data on the phone nor send them to the cloud. The app records the amplitude level of noise at $20Hz$ for periods of $5, 8, 10$ seconds, at intervals of, on average, 30 minutes. Like for location, we extracted high sensor coverage data – we found at least one high sensor coverage day for $3,552$ users. We averaged the amplitude samples during each interval and converted them to decibels (dB) to give an indicative measure of the ambient noise.

Figures 5a, 5b show the average, standard deviation, and median noise during each hour of the day for week days and weekends,



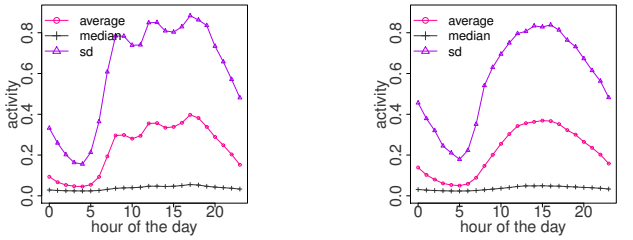
(a) CCDF of locations visited (b) Time spent at locations (c) Location changes (week days) (d) Location changes (weekends)

Figure 4: Location patterns of users



(a) Week days (b) Weekends

Figure 5: Mic sensor: Noise level vs. time of day



(a) Week days (b) Weekends

Figure 6: Accelerometer: Activity level vs. time of day

respectively. Looking at the average, the difference between daytime and nighttime noise level is remarkable, i.e., quieter during night, both during week days and weekends. The increase in noise level starts early (around 6-7am) in week days, but a little later in weekends (from 7-8am onwards). During week days, local peaks of noise are around 12am-1pm and 5pm that overlap with lunch and transit times, perhaps when people are traveling or socializing. The median value follows a similar pattern. This, and the small and quasi-constant standard deviation, indicates small difference between some users and others in terms of their environment.

Lower noise levels at night especially when the user is at home might indicate that he is sleeping. Similarly, change of noise level from low to high in the mornings can be a cue that the user woke up from sleep. Studies have shown that sleep patterns are related to mental diseases [19]. Therefore, our findings on extracting routines using the microphone sensor are an important step in understanding the link between passive sensing data and users' well-being.

3.2 Activity

We used accelerometer data to provide initial insights into the activity level of the user throughout the day. We extracted high sensor coverage days for each of the participants – we found at least one high sensor coverage day for $3,177$ users. Accelerometer data samples consist of $[x, y, z](m/s^2)$ axes data for periods of 5,8,10 seconds, at intervals of, on average, 15 minutes. To estimate the user's activity level, we used the standard deviation of the magnitude of acceleration ($\sqrt{x^2 + y^2 + z^2}$), as it is reported to be effective for activity recognition [46].

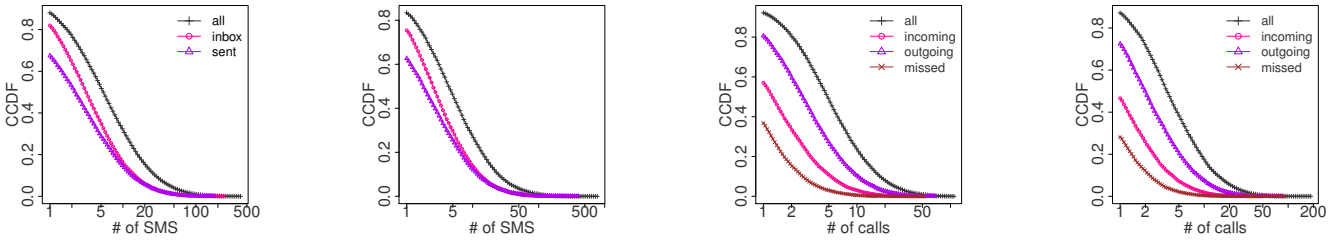
Figures 6a, 6b show the average, standard deviation, and median of participants' activity during each hour of the day, for week days and weekends, respectively. We observe local peaks of average activity around 9am, 12am-1pm and 5pm during week days, i.e., when users transit between home and work, and during lunch time. During weekends, there is a single smooth peak that extends for the whole afternoon and even beyond. The standard deviation follows a similar pattern as the average, which indicates that differences between users (and even between different days for an individual user) are higher during day hours and especially during "common peaks of activity". Median values per hour are almost constant and low, which indicates that our participants are not particularly active.

Several studies [20, 44] have reported that physical activity impacts mental and physical well-being. The extracted longitudinal activity level routines are useful measures for comprehending the relation between the users' physical activity level and well-being.

3.3 Sociability

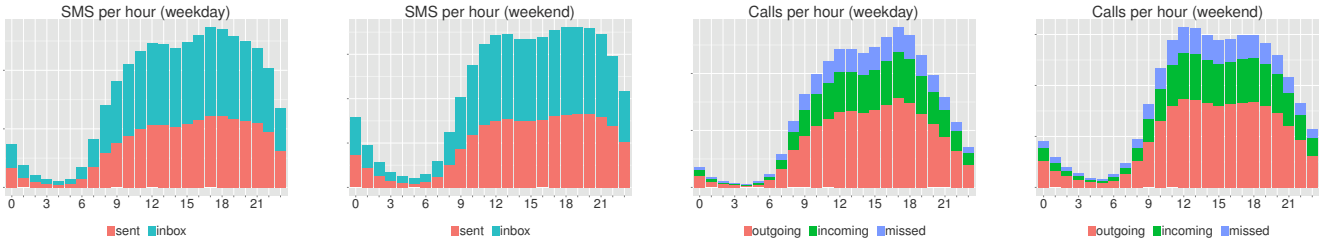
In this subsection we explore the sociability of users, another important factor that can potentially impact users' mental health. We rely on interactions through mobile devices to extract sociability patterns, more specifically, on calls and text messages (SMSs). To preserve their privacy, we do not store SMS content and we *hash* phone numbers using *SHA-256*, a secure one-way hash algorithm. The dataset contains $11,585$ users who have sent/received at least one SMS and $11,465$ who have made/received at least one call.

Distribution of SMSs, Calls: On average, our users send less messages than receive: 6.20 SMSs received, $sd = 10.85$ versus 5.40 sent, $sd = 10.74$. Like in the case of activity, the high values of standard deviation indicate that users exhibit diverse phone usage patterns. That is, some users text a lot, while others barely do. Figures 8a, 8b show the distribution of SMSs received or sent per user per day for week days and weekends. We observe here that users text more during week days than during weekends, al-



(a) Week days (b) Weekends (c) Week days (d) Weekends

Figure 7: CCDF of number of SMS messages and phone calls per hour



(a) Week days (b) Weekends (c) Week days (d) Weekends

Figure 8: SMS messages, Phone calls vs. time of day

though again, there is a large difference across users. With respect to phone calls, our participants make more calls than receive: 2.12 calls received, $sd = 2.93$ versus 4.21 made, $sd = 5.31$. If we also consider *missed* calls, then this difference is reduced: 3.26 incoming vs. 4.21 outgoing. In this case too, the high standard deviation points to the wide variation among users, i.e., some users call a lot, while others hardly do. Figures 8c, 8d show the distribution of calls per user per day for week days and weekends. As with SMSs, users call more during week days than during weekends.

SMSs, Phone calls vs. time of day: Figures 7a, 7b show the distribution of SMS exchanges with respect to time of day for week days and weekends, respectively. We observe *high reciprocity* between the patterns of sent and received SMSs. The distribution of *incoming*, *ongoing* and *missed* phone calls per hour for week days and weekends are shown in Figures 7c, 7d. Although there is almost no difference between the patterns of the different types of calls, there are slight differences between peak hours of calls during week days and weekends: we observe that even though most of the *incoming* and *ongoing* calls take place during daily hours, they are more prone to take place during the evenings on week days, but during the afternoons in weekends. *Missed* calls follow a similar pattern, though the peak hours of missed calls during weekends in the afternoon are more accentuated than in the other kind of calls.

Some of these observations are in-line with findings from existing work, for instance, in [17] the authors reported that their study participants accessed SMS 11.2 times per day and had 5.7 phone calls per day. Further, since sociability (as measured through call patterns, and SMS message exchanges) can be tied to many psychological factors such as loneliness, we later explore if these interaction routines can be useful in tracking users' well-being.

4. RELATING ROUTINES TO PSYCHOLOGICAL PROFILES

4.1 Surveys results

Participants were requested to complete profile-related surveys. These surveys cover a broad range of topics: demographics, personality, gratitude, health, sociability, job satisfaction, life aspirations and connectedness, and the questions were answered using sliding scales (Likert scales). A majority of participants answered the demographics survey, indicating gender, age, race, education and occupational status. However, less participants filled other surveys, where the number of respondents ranged from $3,354$ in the case of personality to $2,420$ in the case of connectedness. Figure 9 contains the distribution of users according to their answers in each survey. This gives an aggregated view of the demographics and life views of the participants.

Our dataset contains slightly more males than females, most of them born between 1970 and 2000 and white. These participants are mainly educated people in full time employment and university students. Results from the Big Five personality traits test [23] revealed that most of our participants do not consider themselves especially agreeable, neither disagreeable (warm, sympathetic), and most of them did not define themselves particularly spontaneous neither self-disciplined, though there is slightly higher presence of more spontaneous participants than self-disciplined. We found that most of the respondents consider themselves very open to new experiences, but not many of them reported to be neither extraverted nor emotionally stable. Most of the participants reported feeling highly grateful towards others and rarely or sometimes connected to other people, and perceiving their health as good. In terms of social relationships, people reported different levels of satisfaction with their families and friends, but bimodal answers when reporting satisfaction with their social partner. People also reported to be

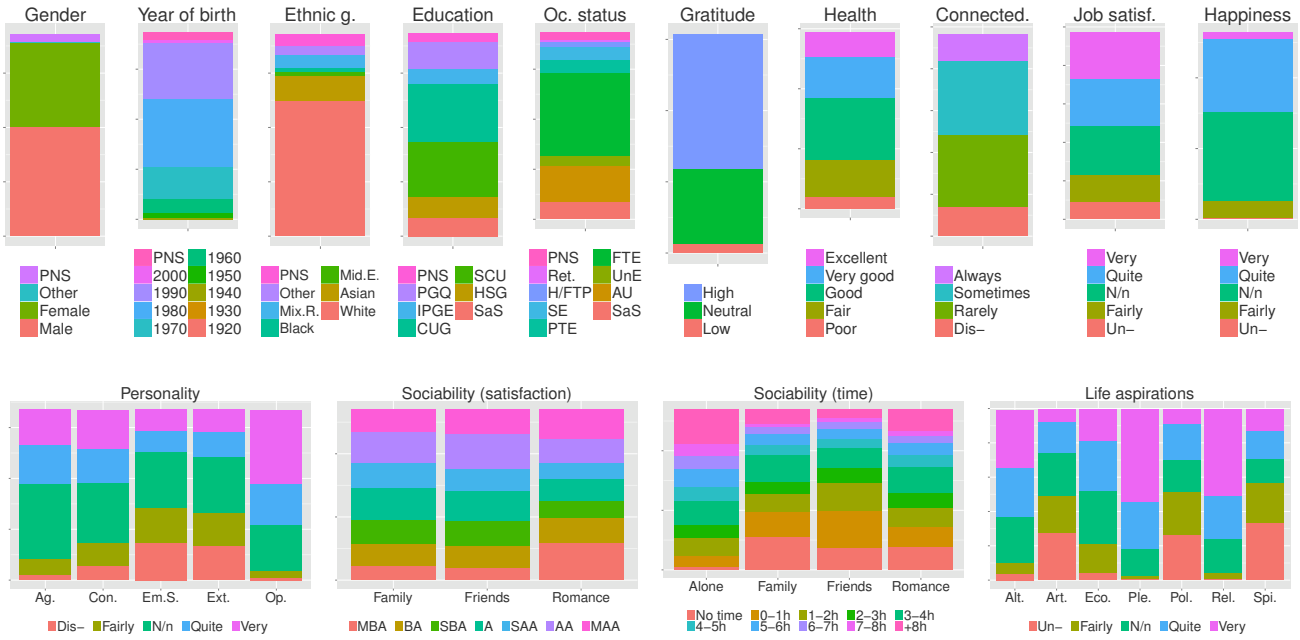


Figure 9: Distribution of users according to their survey answers

most of the time alone, spending similar time with family and partner, but less time, in general, with friends. Participants manifested to be, in general, quite satisfied with their job. Most of the respondents aspire to be altruistic, have fun and have successful social relationships. In terms of economical aspirations, most of the participants reported that for them this was (i) neither important nor unimportant or (ii) quite important. And, in general, the less important aspirations in life for our participants are the political and artistic ones and, overall, the spiritual ones. Finally, we computed the happiness score proposed by Lathia et al. [33] for each user who responded at least once to the affect grid and the survey about their *satisfaction with life scale (SWLS)*, and rated their positive (PA) and negative (NA) at least once. We created a z-score for each of the four components, then added together the z-scores for the grid valence (x-axis on the affect grid), PA and SWLS, and subtracted the z-score for NA. Figure 9 (top-right plot) shows the division of our users into 5 equally-sized groups of happiness level.

Our survey results are useful for the research community to understand the link between various user perceived personality traits and their variance. In the following subsection, we use them to understand the relation between the inferred routines by passive sensing data, and the demographics and life views of the participants.

4.2 Relating sensing activity and profile

In this subsection, we explore *do people with similar profiles have similar routines as sensed by phone sensors?* To answer this question we use the Kruskal-Wallis test [28] that tests the null hypothesis that samples in two or more groups are drawn from populations with the same median values. We applied this non-parametric alternative to one-way ANOVA [38] (ANOVA could not be applied due to the lack of normality in the distribution of users' responses) to determine if participants with different demographics and live views have also different *median values* of sensed routine.

For each profile-related survey, we consider all the participants for whom we have at least one high sensor coverage day. We obtain their routines by computing the average values of the different parameters sensed during different times of the day (those in which users' are more free to decide where to stay, what to do, etc., i.e., non-working hours). These are the standard deviation of the magnitude of acceleration, the amplitude of the noise sensed with the microphone, the number of calls recorded and the number of texts messages exchanged. For the mobility data, we consider the average number of different locations visited per day.

Figure 10 shows the results of applying the Kruskal-Wallis test to check the relation between participants' demographics, personality, etc., and their sensed data at different times of the day. Different colors are used to represent the distinct types of sensed data (accelerometer in green, microphone in blue, SMSs in purple and calls in turquoise). With this, each position $(x, y(t))$ in the table is colored in dark or light depending on if the p-value in the Kruskal Wallis test is lower or higher than 0.05 (the significance level) and, therefore, the null hypothesis can or cannot be rejected. That is, position $(x, y(t))$ is colored in dark if the median of the sensed values with the (y) sensor at the time of the day (t) for those users that answered similarly to the question (x) in the survey, is different from the median values of users who provided a different answer to this question. Broadly speaking, position $(x, y(t))$ is colored in dark if there exist relationship between users' profiles and their location, activity, environment, etc. sensed with their mobile devices.

4.2.1 Findings

Results show that almost all participants' demographic parameters (gender, year of birth, occupational status, education, and ethnic group) are related with their sensed data. That is, *routines extracted by the user's sensor data can be used to understand the demographics of users*. Looking at the median values of activity and noise level, we observed that males tend to be more active than females (0.27 vs 0.21), and staying in noisier places (52.39 vs 50.26 dBs). However, females send/receive more text messages per day (6.77 vs the 4.98 of the males), though their call patterns are similar (3.28 vs 3.43 calls). In terms of age, the activity, the level of noise, number of messages and calls increase with the age for those users

that were born before 2000. For teenagers, this tendency is inverted for all the parameters considered. In the case of location, the ones that visit more locations per day during weekdays are those who were born between 1970 and 1990. For the occupational status, we see that retired participants, homemakers and unemployed are the ones less active (median values per day of 0.11, 0.17 and 0.20, respectively), whereas those at university and on full-time/self employment are usually the most active (0.25, 0.24 and 0.23). The latter, together with those in part-time employment, are also the ones that visit more locations per day. Note that these values are obtained from sensed data and might be affected by the fact that some participants are most likely to carry their phones than others. For instance, homemakers usually do not carry their phone at home, though they might be moving around the house.

Results from the non-demographic variables, looking again at median values of sensor data, show that disagreeable people tend to text more and call less than more agreeable (7.74 messages –2.09 calls– per day for disagreeable people vs 6.61 messages –3.11 calls– for the very agreeable ones). Also, the most emotional stable participants tend to be more active, stay in more noisy places and text less than unstable participants. And more extraverted participants are more active during evenings and weekdays, stay in louder places and make/receive more calls than less extraverted ones. Gratitude is also related with the messages and calls patterns behavior, being the most grateful users the ones that communicate more through their phones. In terms of health, participants that reported feeling in better health are more active and less prone to use text messages. Regarding social relationships, participants that (i) have a romantic partner, (ii) are more satisfied with their social relationships or (ii) spend more time with other people tend to communicate more through calls than the rest. We found that the respondents more satisfied with their job tend to visit more locations per day during weekdays than the rest. In terms of life aspirations, we found that participants that pursue to have fun and an exciting lifestyle (pleasure) are the most active –as they might be working towards this goal, especially in the evenings. The number of calls also is positively related with the spiritual, relationships, economic and political aspirations. Finally, the more connected the participants, the more active they are, and the more calls they do/receive. Finally, we found that happiness is positively related with activity. That is, happy users live active lives.

In summary, we found that, in the majority of cases, users with the same demographics have similar sensor data patterns. Further, in some cases, the latter is also related with various user perceived psychological factors. *We believe that these findings will be helpful for the research community to understand users' demographics through routines extracted from their sensor data.*

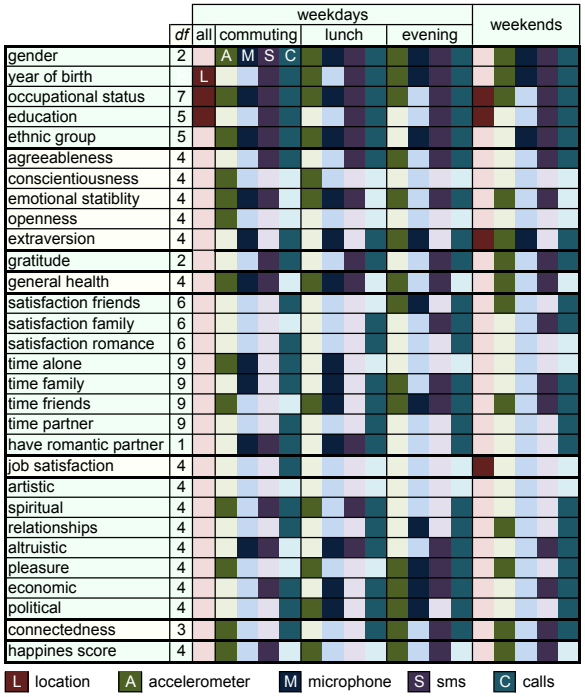


Figure 10: Correlation measures between personality tests and mobile sensed data.

5. INFERRING MOOD FROM SENSOR DATA

In this section we study the predictability of users' mood by using smartphone sensor data. Specifically, we investigate the possibility of assessing user's momentary mood at time *t* from his environment (microphone), activity (accelerometer), and sociability (messages and phone calls) data sensed before and after *t*. We then explore the correlation of sensor data with the self-reported mood.

5.1 Data preprocessing

If the user reported his mood N times at t_m, $m \in \{1, N\}$, our aim is to infer user's mood ($x(t_m) = positive, negative$, $y(t_m) = alert, sleepy$) at t_m from behavioral data collected with his smartphone sensors before and after the self report. We characterize his previous behavior as the *difference* between his data sensed in the interval between the current and the prior self report of that day ($[t_{m-1}, t_m]$) –or the zero hours of that day in the case of $m = 1$ ($[0, t_m]$), and his average data sensed during the same interval in all his *high sensor coverage* days. Reciprocally, we compute his future behavior as the *difference* between his average and his sensed data in the interval between the current and the subsequent self report of that day ($[t_m, t_{m+1}]$) –or the 24 hours of that day in the case of $m = N$ ($[t_m, 24]$). For the accelerometer and microphone sensors, we apply *Dynamic Time Warping (DTW)* [8] to obtain the distance (difference) between the sensor readings each day during the observation intervals and the average readings during those intervals for the given user. *DTW* finds an optimal alignment between two time-dependent sequences S_1 and S_2 by warping the time dimension in S_1 that minimizes the difference between the two series so that time series do not need to be of equal length. For text messages and phone calls, we consider the difference between the number of messages (phone calls) exchanged each day during the observation intervals and the average number of them exchanged during those intervals for the given user. Further, we observed that users report their mood differently: while some users use most of the values in the x and y axes of the *affect grid*, others reported their mood using only a small portion. Therefore, we consider that the user is in a positive (alert) mood if he reports a x-value (y-value) higher than the median of the x-values (y-values) reported by him. Analogously, he is in a negative (sleepy) mood if this value is lower than the median.

5.2 Experimental setup

We used a Deep Neural Network (DNN) of stacked Restricted Boltzmann Machines (RBMs) –a type of Markov Random Field that include visible and hidden units, for mood classification. Layers of the network are formed using multiple RBMs stacked together where the hidden units from one set of RBMs act as the visible layer for the next. A hidden unit (k) computes its own state (y_k) –that is passed on to subsequent layers– in two stages. First, it

computes an intermediate state (x_k) using $b_k + \sum_i y_i w_{ik}$, where b_k is a unit-specific bias term, y_i is the state of each unit in the prior layer, and w_{ik} is the weight between unit k and again each prior layer unit. Second, it applies an activation function to x_k –we used rectified linear functions [39] in which case $y_k = max(0, x_k)$. We construct the input layer using users' behavior sensed before and after the self report. Because of the sensitivity of later deep learning stages to feature scaling [25] we normalize all distances/differences to have zero mean and unit standard deviation. Throughout the remaining stacked RBMs in the network we use ReLU units. The exception to this occurs with the output layer where each unit corresponds to a mood inference class, so that unit states can be interpreted as posterior probabilities.

We used 8 features (corresponding to users' sensed behaviors before and after the mood self-report, which were obtained using accelerometer, microphone, texts and calls data) and 2 classes (positive and negative –alert and sleepy– mood) in our classification model. We set up a DNN with 5 layers, including 3 hidden layers with $1,024$ nodes per layer, which led to use $3,082$ units, resulting in over $2.1M$ parameters to be determined during training. The training and testing was performed with 5-fold cross validation. All training procedures are implemented in python using the *Theano* deep learning library [6]. We used grid search to determine the values of the hyper-parameters, setting the learning rate to 0.05, training epochs to 100, and batch size to 10.

For this study we considered users for whom we have at least one *high sensor coverage* day for accelerometer, microphone, messages and phone calls, and who reported their mood at least once during that day. Our sample contains $10,182$ mood self-reports from 726 users, of which $7,779$ reports were completed during week days by 650 users and $2,403$ during weekends by 431 users.

5.3 Results

Figure 11a shows the accuracy (per mood self report) obtained when considering all days, only week days, and only weekends. In spite of several constraints such as (i) users' mood might depend on many external factors that are not being monitored (e.g. an email received or a conversation with a friend), (ii) the noise in the large-scale dataset and (iii) the diversity and variance among users, we achieve 68% accuracy when classifying users' positive/negative mood (x) during weekends. We observe that the accuracy is higher for inferring x from sensing data than y, and when considering self reports during weekends. This could be because the weekend behavior of users is generally more dependent on their mood than weekdays, therefore, the differences between their routines during weekends with respect to their averages is typically higher.

The classification accuracy observed in our evaluation is limited by the noise introduced by the scale of the uncontrolled study. This is in-line with existing studies [31] that showed that classification accuracy lowers when considering a large scale and diverse set of participants. Although the average accuracy per user approximates the one per self report, we found high differences between groups of users: these differences are likely due to users' diversity and possibly low number of observations for some users.

5.4 Informativeness of each sensor

Now we perform a deeper analysis to explore the relation between passively collected sensor data and mood. For each user, we compute the Pearson correlation between each feature described in the previous section, i.e. difference between user's sensed data before and after the self report with respect to his average sensed data during this time, and his reported mood (valence –x– and arousal –y– in the *affect grid*) during his *high sensor coverage* days.

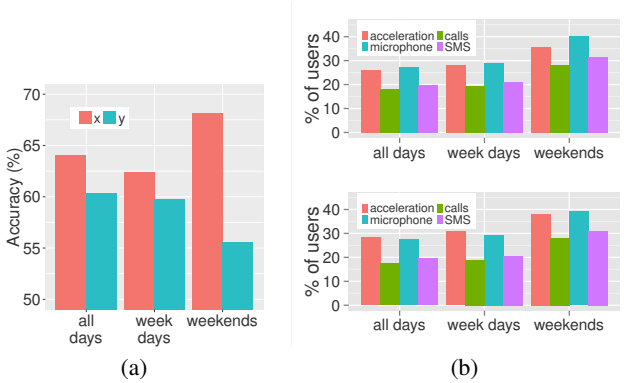


Figure 11: (a) Accuracy of the classification for valence –*x*– and arousal –*y*– in the affect grid. (b) Percentage of users for whom the absolute value of the Pearson correlation between passive sensing data and mood (*x* in the top plot, *y* in the bottom) is higher than 0.5.

For this analysis, our sample contains mood self-reports and sensing data from $1,556$ users for accelerometer analysis, $1,656$ for microphone, $5,469$ for text messages and $8,247$ for phone calls respectively. Figure11b shows the percentage of users for whom the absolute value of the Pearson correlation between his behavior sensed before or after the self-report and the self-reported mood (valence –x– in the top plot, arousal –y– in the bottom) was significative (higher than 0.5). We found significative correlation for a higher percentage of the users when considering data sensed with the microphone and the accelerometer than when considering text messages and phone calls records. That is, *the level of users' activity and noise level in their environment are more informative of users' mood than their sociability level.* Also, and in-line with the results in previous subsection, the percentage of users for whom the correlation is significative is higher during weekends than on weekdays, when users have more leisure time. Regarding valence and arousal, we have not found significative differences that justify the different accuracy between classifiers reported in previous section.

6. RELATED WORK

Interest in using smartphones for mental well-being has been increasing at a brisk pace [21, 30, 34]. We broadly divide existing work at the intersection of mobile sensing and mental health into the following categories.

Mobile experience sampling. The advent of smartphones has spurred researchers to devise innovative experience sampling techniques that could one day completely replace traditional pen and paper methods. Prominent among these are MyExperience [21], Momento [11], Context-aware Experience Sampling Tool [26], MIT Funf [7], and Aware Framework [18]. Our work draws inspiration from these on how to blend-in user initiated self-reports, system triggered experience sampling, and passive sensor data collection. Our focus is on mood detection, unlike most of these general tools. In this paper we describe our mood reporting interface which makes use of an affect grid [47] and concentrate on the analysis of our large scale dataset in order to give insights into the relationship of mood and user personality, demographics and passive sensing.

Affect recognition. Emotion, mood, and stress detection has been the focal point of existing work in affect recognition. EmotionSense [45] and StressSense [36] are smartphone systems that

use acoustic signals extracted from the microphone sensor of user's phone to detect emotion, stress, respectively. MoodScope [34] takes an orthogonal approach and uses smartphone usage patterns, for example, browsing history, phone calls, to infer the user's mood. MIT Mood meter [24] uses a computer vision technique to detect smiles in a large community using cameras installed at various locations. While a part of our work too dwells in the field of mood inference, our primary contribution is to fathom if a diverse set of sensor signals can be used in estimating user's mood using a large scale deployment. Further, unlike us, most of these studies used a limited number of participants ($14 - 32$ users).

Behavior monitoring. Studying human behavior using mobile devices has been an active topic of much recent research [10, 12, 30, 50]. Although a substantial percentage of existing work focussed on physical activity [14, 12], there has been growing interest in tracking diverse set of signals that influence well-being (Bewell [30], StudentLife [50]). Some systems, like ours, combine passive sensing with experience sampling [2], while others solely rely on manual input from users [29]. Unlike these systems, our work analyzes a much larger passive sensing and self-report dataset by considering a diverse set of signals such as noise, location, movement, and communication patterns. Further, we also explore the relation between users' demographics, personality, and their sensor data. Existing work that used purpose-built devices to monitor user behavior include Electronically Activated Reader (EAR) [37], Mobile Sensing Platform (MSP) [12], and Sociometer [13]. Our work instead focuses on off-the-shelf mobile devices and in-the-wild data collection using mobile application stores.

Large-scale studies using mobile devices. Our work has been inspired by many large-scale studies using smartphones such as Device Analyzer (phone usage data) [49], Energy Emulation Toolkit study [42], and app usage studies [9, 22] – each of these used data from thousands of users. Most existing work at the crossroads of smartphone sensing and mental health used a limited number of users in their studies, usually less than 100 participants [36, 34, 30]. We have successfully deployed our application on the *Google Play* store and collected data from thousands of users to understand the feasibility of estimating well-being from smartphone sensor data.

7. CONCLUSIONS AND FUTURE WORK

Our application for Android was conceived as a tool for mood monitoring that combines mobile sensing with self-reports from users. By collecting data from sensors in the user's phone (microphone, accelerometer, location, messages, and calls) at different sampling rates, and asking users to report their mood multiple times a day, we give simple feedback to users about their daily variation of mental state in relation to the sensed environment or activity. The analysis of the large volume of data collected for three years shed some light on (i) the utility of mobile sensing in identifying regularities in users' routines and (ii) the underlying relation between users' activity, sociability or mobility with respect to their demographics and different psychological aspects, such as perception of health, life satisfaction, and connectedness. Moreover, the longitudinal analysis of the sensed data combined with self-reports revealed the relation between users' routines and reported mood. We report results from an analysis $10x$ bigger than similar studies and still maintain high accuracy.

In addition to these findings, our work provides developers of future mobile sensing applications with useful insights on the utility of passive sensing for behavior monitoring, specifically, for mood monitoring. In the future, we plan to use our passive sensing findings to provide users with better feedback and, ultimately, better behavior interventions to improve their mental health and well-being.

8. ACKNOWLEDGMENTS

This work was supported by the EPSRC through Grants UBHAVE (EP/I032673/1) and GALE (EP/K019392). The authors are grateful to Petko Georgiev for his assistance with the deep learning classifier.

9. REFERENCES

[1] Australian bureau of statistics (2009). **http://www.abs.gov.au/ausstats/abs@.nsf/mf/4326.0**.

[2] Mappiness. **http://www.mappiness.org.uk**.

[3] Mental disorders - who. **http://www.who.int/mediacentre/factsheets/fs396/en/**.

[4] Mental health uk. **https://www.mentalhealth.org.uk/statistics/mental-health-statistics-uk-and-worldwide**.

[5] Nih - any mental illness (ami) among u.s. adults. **http://www.nimh.nih.gov/health/statistics/prevalence/any-mental-illness-ami-among-us-adults.shtml**.

[6] Theano deep learning. **http://deeplearning.net/software/theano**.

[7] Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput. 7*, 6 (2011).

[8] Bemdt, D. J., and Clifford, J. Using dynamic time warping to find patterns in time series, 1994.

[9] Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *MobileHCI '11*, ACM (2011).

[10] Canzian, L., and Musolesi, M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *UbiComp'15*, ACM (2015), 1293–1304.

[11] Carter, S., Mankoff, J., and Heer, J. Momento: Support for situated ubicomp experimentation. In *CHI '07*, ACM (2007).

[12] Choudhury, T., Borriello, G., Consolvo, S., Haehnel, D., Harrison, B., Hemingway, B., Hightower, J., Klasnja, P. ., Koscher, K., LaMarca, A., Landay, J. A., LeGrand, L., Lester, J., Rahimi, A., Rea, A., and Wyatt, D. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing* (2008).

[13] Choudhury, T., and Pentland, A. Sensing and modeling human networks using the sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, ISWC '03, IEEE Computer Society (2003).

[14] Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., and Landay, J. A. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, ACM (New York, NY, USA, 2008), 1797–1806.

[15] Csikszentmihalyi, M., and Larson, R. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Springer Netherlands, Dordrecht, 2014, ch. Validity and Reliability of the Experience-Sampling Method.

[16] Diener, E., and Seligman, M. E. Beyond money: Toward an economy of well-being. *Psychological Science in the Public Interest 5*, 1 (2004), 1–31.

[17] Do, T. M. T., Blom, J., and Gatica-Perez, D. Smartphone usage in the wild: A large-scale analysis of applications and context.

[18] Ferreira, D., Kostakos, V., and Dey, A. Aware: mobile context instrumentation framework. *Frontiers in ICT 2* (2015).

[19] Foley, D., Ancoli-Israel, S., Britz, P., and Walsh, J. Sleep disturbances and chronic disease in older adults: Results of the 2003 national sleep foundation sleep in america survey. *Journal of Psychosomatic Research 56*, 5 (2004), 497 – 502.

[20] Fox, K. R. The influence of physical activity on mental well-being. *Public Health Nutrition 2* (3 1999), 411–418.

[21] Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. A. Myexperience: A system for in situ tracing and capturing of user feedback on mobile phones. In *MobiSys '07*, ACM (2007).

[22] Girardello, A., and Michahelles, F. Appaware: Which mobile applications are hot? In *MobileHCI '10*, ACM (2010).

[23] Goldberg, L. R. The development of markers for the big-five factor structure. *Psychological Assessment* (1992).

[24] Hernandez, J., Hoque, M. E., Drevo, W., and Picard, R. W. Mood meter: Counting smiles in the wild. In *UbiComp '12*, ACM (2012).

[25] Hinton, G. A practical guide to training restricted boltzmann machines. *Momentum 9*, 1 (2010), 926.

[26] Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., and Bao, L. A context-aware experience sampling tool. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ACM (2003).

[27] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. Identifying important places in people's lives from cellular network data. In *Pervasive'11* (2011).

[28] Kruskal, W. H., and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association 47*, 260 (1952), 583–621.

[29] Lamminmaki, E., Parkka, J., Hermersdorf, M., Kaasinen, J., Samposalo, K., Vainio, J., Kolari, J., Kulju, M., Lappalainen, R., and Korhonen, I. Wellness diary for mobile phones. In *In Proc. EMBEC '05.* (2005).

[30] Lane, N. D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., and Campbell, A. T. Bewell: A smartphone application to monitor, model and promote wellbeing. In *Pervasive Computing Technologies for Healthcare* (2011).

[31] Lane, N. D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, A. T., and Zhao, F. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *UbiComp'11*, ACM (2011), 355–364.

[32] Lathia, N., Rachuri, K., Mascolo, C., and Roussos, G. Open source smartphone libraries for computational social science. In *UbiComp '13 Adjunct*, ACM (2013).

[33] Lathia, N., Sandstrom, G. M., Mascolo, C., and Rentfrow, P. J. Happier people live more active lives: Using smartphones to link happiness and physical activity. *PLoS ONE* (2017).

[34] LiKamWa, R., Liu, Y., Lane, N. D., and Zhong, L. Moodscope: Building a mood sensor from smartphone usage patterns. In *MobiSys '13*, ACM (2013).

[35] Liu, P., Tov, W., Kosinski, M., Stillwell, D. J., and Qiu, L. Do facebook status updates reflect subjective well-being? *Cyberpsychology, Behavior, and Social Networking* (2015).

[36] Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *UbiComp '12*, ACM (2012).

[37] Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., and Price, J. H. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers 33*, 4 (2001), 517–523.

[38] Miller Jr, R. G. *Beyond ANOVA: basics of applied statistics*. CRC Press, 1997.

[39] Nair, V., and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (2010), 807–814.

[40] Napa Scollon, C., Prieto, C.-K., and Diener, E. *Assessing Well-Being: The Collected Works of Ed Diener*. Springer Netherlands, Dordrecht, 2009, ch. Experience Sampling: Promises and Pitfalls, Strength and Weaknesses, 157–180.

[41] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. An empirical study of geographic user activity patterns in foursquare. In *Proc. of the 5th Int'l AAAI Conference on Weblogs and Social Media* (2011).

[42] Oliver, E. A., and Keshav, S. An empirical approach to smartphone energy level prediction. In *UbiComp '11*, ACM (2011).

[43] Parate, A., Chiu, M.-C., Chadowitz, C., Ganesan, D., and Kalogerakis, E. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *MobiSys '14*, ACM (2014).

[44] Penedo, F. J., and Dahn, J. R. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Current Opinion in Psychiatry 18*, 2 (2005), 189–193.

[45] Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *UbiComp '10*, ACM (2010).

[46] Ravi, N., Dandekar, N., Mysore, P., and Littman, M. L. Activity recognition from accelerometer data. In *IAAI'05* (2005).

[47] Russell, J. A., Weiss, A., and Mendelsohn, G. A. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology* (1989).

[48] Sandstrom, G. M., Lathia, N., Mascolo, C., and Rentfrow, P. J. Putting mood in context: Using smartphones to examine how people feel in different locations. *Journal of Research in Personality* (2016).

[49] Wagner, D. T., Rice, A., and Beresford, A. R. *Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer International Publishing, 2014, ch. Device Analyzer: Understanding Smartphone Usage.

[50] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp '14*, ACM (2014).