

StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell

Abstract Much of the stress and strain of student life remains hidden. The *StudentLife* continuous sensing app assesses the day-to-day and week-by-week impact of workload on stress, sleep, activity, mood, sociability, mental well-being and academic performance of a single class of 48 students across a 10 weeks term at Dartmouth College using Android phones. Results from the StudentLife study show a number of significant correlations between the automatic objective sensor data from smartphones and mental health and educational outcomes of the student body. We propose a simple model based on linear regression with lasso regularization that can accurately predict cumulative GPA. We also identify a Dartmouth term lifecycle in the data that shows students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while positive affect, sleep, conversation and activity drops off. The StudentLife dataset is publicly available on the web.

R. Wang (✉) • F. Chen • Z. Chen • T. Li • X. Zhou • A.T. Campbell
Dartmouth College, Hanover, NH, USA
e-mail: ruiwang@cs.dartmouth.edu; chentc@cs.dartmouth.edu; zhenyu@cs.dartmouth.edu;
ltx@cs.dartmouth.edu; xia@cs.dartmouth.edu; campbell@cs.dartmouth.edu

G. Harari
The University of Texas at Austin, Austin, TX, USA
e-mail: dror.ben-zeev@dartmouth.edu

S. Tignor
Northeastern University, Boston, MA, USA
e-mail: gabriella.harari@utexas.edu

D. Ben-Zeev
Department of Psychiatry & Behavioral Sciences, University of Washington, 1959 NE Pacific
Street, Box 356560, Room BB1644, WA 98195-6560, Seattle
e-mail: tignor.s@husky.neu.edu

Introduction

Many questions arise when we think about the academic performance of college students. Why do some students do better than others? Under similar conditions, why do some individuals excel while others fail? Why do students burnout, drop classes, even drop out of college? What is the impact of stress, mood, workload, sociability, sleep and mental well-being on educational performance? In this paper, we use smartphones carried by students to find answers to some of these pressing questions.

Consider students at Dartmouth College, an Ivy League college in a small New England college town. Students typically take three classes over a 10-week term and live on campus. Dartmouth classes are generally demanding where student assessment is primarily based on class assignments, projects, midterms and final exams. Students live, work and socialize on a small self-contained campus representing a tightly-knit community. The pace of the 10 week Dartmouth term is fast in comparison to a 15 week semester. The atmosphere among the students on campus seems to visibly change from a relaxed start of term, to an intense midterm and end of term. Typically classes at Dartmouth are small (e.g., 25–50 students), but introductory classes are larger (e.g., 100–170), making it difficult for a faculty to follow the engagement or performance of students on an individual level. Unless students contact a student dean or faculty about problems in their lives, the impact of such challenges on performance remains hidden.

To shine a light on student life we develop the *StudentLife* [51] smartphone app and sensing system to automatically infer human behavior in an energy-efficient manner. The *StudentLife* app integrates MobileEMA, a flexible ecological momentary assessment [45] (EMA) component to probe students' states (e.g., stress, mood) across the term. We administer a number of well-known pre-post health and behavioral surveys at the start and end of term. We present the results from a deployment of *StudentLife* on Google Nexus 4 Android phones at Dartmouth College.

StudentLife makes a number of contributions. *First*, to the best of our knowledge we are the first to use automatic and continuous smartphone sensing to assess mental health, academic performance and behavioral trends of a student body. *Second*, we identify strong correlation between automatic sensing data and a broad set of well-known mental well-being measures, specifically, PHQ-9 depression, perceived stress (PSS), flourishing, and loneliness scales. Results indicate that automatically sensed conversation, activity, mobility, and sleep have significant correlations with mental well-being outcomes. we propose for the first time a model that can predict a student's cumulative GPA using automatic behavioral sensing data from smartphones. We use the *Lasso* (Least Absolute Shrinkage and Selection Operator) [48] regularized linear regression model as our predictive model. Our prediction model indicates that students with better grades are more conscientious, study more, experience positive moods across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period, and experience change in their conversation duration patterns later in the term. *Third*, we observe trends

in the sensing data, termed the *Dartmouth term lifecycle*, where students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while activity, sleep, conversation, positive affect, visits to the gym and class attendance drop.

Related Work

There is a growing interest in using smartphone sensing [9, 10, 12, 13, 17, 38, 50] to infer human dynamics and behavioral health [8, 20, 24, 26, 29, 34, 36, 41, 42]. The StudentLife study is influenced by a number of important behavioral studies: (1) the friends-and-families study [8], which uses Funf [4] to collect data from 130 adult members (i.e., post-docs, university employees) of a young family community to study fitness intervention and social incentives; and (2) the reality mining project [22], which uses sensor data from mobile phones to study human social behavior in a group of students at MIT. The authors show that call records, cellular-tower IDs, and Bluetooth proximity logs accurately detect social networks and daily activity.

There is little work on correlations between continuous and automatic sensing data from smartphones and mental health outcomes such as PHQ-9. However, the authors in [41] use wearable sensors (i.e., Intel's mobile sensing platform) to study the physical and mental well-being of a group of 8 seniors living in a continuing care retirement community over a single week. The retirement community study [41] is the first to find correlations with depression and continuous sensing measures from wearables. In [40], the authors monitor bipolar disorder in patients using wearable sensors, but the project does not enable continuous sensing data. In [11, 25], the authors present an approach that collects self-assessment and sensor data on a smartphone as a means to study patients' mood. They find that self-reported activity, stress, sleep and phone usage are strongly correlated with self-reported mood. Health Buddy [28] asks patients a series of questions about symptoms of depression to help mental health service providers monitoring patients' symptoms. No continuously sensing is used. Mobilyze is an intervention system [14] that uses smartphones to predict self-reported states (e.g., location, alone, mood) using machine learners. Results indicate that Mobilyze can predict categorical contextual states (e.g., location, with friends) with good accuracy but predicting internal states such as mood show poorer predictive power.

There is a considerable interest in studying the health and performance of students. However, no study has used smartphone sensing to study these issues. In [49], the authors study the effect of behaviors (i.e., social support, sleep habits, working hours) on grade points based on 200 randomly chosen students living on the campus at a large private university. However, this study uses retrospective survey data manually entered by users to assess health and performance. Watanabe [53, 54] uses a wearable sensor device to investigate the correlation between face-to-face interaction between students during break times and scholastic performance. Previous research [23] aimed at predicting performance has used a neural network

model to predict student's grades from their placement test scores. Various data collected from entering students are used in [37] to predict student academic success using discriminant function analysis. Kotsiantis and Pintelas [31] proposes a regression model to predict the student's performance from their demographic information and tutor's records. Romero et al. [43] applies web usage mining in e-learning systems to predict students' grades in the final exam of a course. In [57], the authors propose an approach based on multiple instance learning to predict student's performance in an e-learning environment. Recent work [46] showed that they can predict a student is at risk of getting poor assessment performance using longitudinal data such as previous test performance and course history. To the best of our knowledge there is no work on using passive sensor data from smartphones as a predictor on academic success.

Study Design

In this section, we discuss how participants were recruited from the student body, and then describe our data collection process. We also discuss compliance and data quality issues in this longitudinal study.

Participants

All participants in the study were voluntarily recruited from the CS65 Smartphone Programming class [1], a computer science programming class at Dartmouth College offered to both undergraduate and graduate students during Spring term in 2013. This study is approved by the Institutional Review Board at Dartmouth College. Seventy five students enrolled in the class and 60 participants joined the study. As the term progressed, 7 students dropped out of the study and 5 dropped the class. We remove this data from the dataset analyzed in section "Results". Among the 48 students who complete the study, 30 are undergraduates and 18 graduate students. The class demographics are as follows: 8 seniors, 14 juniors, 6 sophomores, 2 freshmen, 3 Ph.D students, 1 second-year Masters student, and 13 first-year Masters students. In terms of gender, 10 participants are female and 38 are male. In terms of race, 23 participants are Caucasians, 23 Asians and 2 African-Americans. Forty eight participants finished the pre psychological surveys and 41 participants finished all post psychological surveys.

All students enrolled in the class were offered unlocked Android Nexus 4s to complete assignments and class projects. Many students in the study had their own iPhones or Android phones. We denote the students who use their own Android phones to run the StudentLife sensing system as *primary users* and those who use the Nexus 4s as *secondary users*. Secondary users have the burden of carrying both their own phones and the Nexus 4s during the study. We discuss compliance and data quality of users in section "Compliance and Data Quality".

Table 1 Mental well-being and personality surveys

Survey	Measure
Patient health questionnaire (PHQ-9) [32]	Depression level
Perceived stress scale (PSS) [19]	Stress level
Flourishing scale [21]	Flourishing level
UCLA loneliness scale [44]	Loneliness level
Big five inventory (BFI) [27]	Personality traits

Study Procedure

The StudentLife study consists of orientation, data collection and exit stages. In addition, we deployed a number of management scripts and incentive mechanisms to analyze and boost compliance, respectively.

Entry and Exit During the orientation stage, participants sign the consent form to join the study. Each student is given a one-on-one tutorial of the StudentLife system and study. Prior to signing the consent form, we detail the type of data to be collected by the phone. Students are trained to use the app. Students do not need to interact with the background sensing or upload functions. They are shown how to respond to the MobileEMA system. A series of entry health and psychological baseline surveys are administered using SurveyMonkey as discussed in section “Results” and shown in Table 1. As part of the entry survey students provide demographic and information about their spring term classes. All surveys are administered using SurveyMonkey [7]. These surveys are pre measures which cover various aspects of mental and physical health. Outcomes from surveys (e.g., depression scale) are used as ground truth in the analysis. During the exit stage, we administered an exit survey, interview and the same set of behavioral and health surveys given during the orientation stage as post measures.

Data Collection The data collection phase lasted for 10 weeks for the complete spring term. After the orientation session, students carried the phones with them throughout the day. Automatic sensing data is collected without any user interaction and uploaded to the cloud when the phone is being recharged and under WiFi. During the collection phase, students were asked to respond to various EMA questions as they use their phones. This in-situ probing of students at multiple times during the day provides additional state information such as stress, mood, happiness, current events, etc. The EMA reports were provided by a medical doctor and a number of psychologists on the research team. The number of EMAs fired each day varied but on average 8 EMAs per day were administered. For example, on days around assignment deadlines, we scheduled multiple stress EMAs. We set up EMA schedules on a week-by-week basis. On some days we administer the same EMA (e.g., PAM and stress) multiple times per day. On average, we administer 3–13 EMA questions per day (e.g., stress). The specific EMAs are discussed in section “StudentLife Dataset”.

Data Collection Monitoring StudentLife includes a number of management scripts that automatically produce statistics on compliance. Each time we notice students' phones not uploading daily data (e.g., students left phones in their dorms during the day), or gaps in weekly data (e.g., phones powered down at night), or no response to EMAs, we sent emails to students to get them back on track.

Incentives To promote compliance and data quality, we offer a number of incentives across the term. First, all students receive a StudentLife T-shirt. Students could win prizes during the study. At the end of week 3, we gave away 5 Jawbone UPs to the 5 top student collectors randomly selected from the top 15 collectors. We repeated this at week 6. We defined the top collectors as those providing the most automatic sensing and EMA data during the specific period. At the end of the study, we gave 10 Google Nexus 4 phones to 10 collectors who were randomly selected from the top 30 collectors over the complete study period.

Privacy Considerations Participants' privacy is a major concern of our study. In order to protect participants' personal information, we fully anonymize each participant's identity with a random user id and kept the user id map separate from all other project data so that the data cannot be traced back to individuals. Call logs and SMS logs are one-way hashed so that no one can get phone numbers or messages from the data. Participants' data is uploaded using encrypted SSL connections to ensure that their data cannot be intercepted by third-parties. Data is stored on secured servers. When people left the study their data was removed.

Compliance and Data Quality

The StudentLife app does not provide students any feedback by design. We do not want to influence student behavior by feedback, rather, we aim to unobtrusively capture student life. Longitudinal studies such as StudentLife suffer from a drop in student engagement and data quality. While automatic sensor data collection does not introduce any burden other than carrying a phone, collecting EMA data can be a considerable burden. Students typically are compliant in responding to survey questions at the start of a study, but as the novelty effect wears off, student compliance drops.

There is a 60/40 split of iPhone/Android users in the study group. Of the 48 students who completed the study, 11 are primary phone users and 37 secondary users. One concern is that the burden of carrying two phones for 10 weeks would result in poorer data quality from the secondary users compared to the primary users. Figure 1a shows the average hours of sensor data we have collected from each participant during the term. As expected, we observe that primary users are better data sources, but there is no significant difference. We can clearly see the trend of data dropping off as the term winds down. Achieving the best data quality requires 24 h of continuous sensing each day. This means that users carry their phones and power their phones at night. If we detect that a student leaves their phone at the

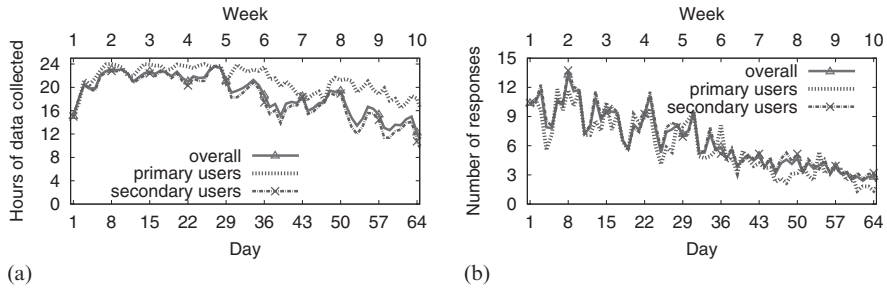


Fig. 1 Compliance and quality of StudentLife data collected across the term. (a) Automatic sensing data quality over the term. (b) EMA data quality over the term

dorm during the day, or it is powered down, then we remove that data from the dataset. The overall compliance of collecting automatic sensing data from primary and secondary users over the term is 87% and 81%, respectively.

Figure 1b shows the average number of EMA responses per day for primary and secondary users. The figure does not capture compliance per se, but it shows that secondary users are slightly more responsive to EMAs than primary users. On average we receive 5.8 and 5.4 EMAs per day per student across the whole term from secondary and primary users, respectively. As the term progresses there is a drop in both administered EMAs and responses. However, even at the end of term, we still receive over 2 EMAs per day per student. Surprisingly, secondary users (72%) have better EMA compliance than primary users (65%). During the exit survey, students favored short PAM-style EMAs (see Fig. 3a), complained about the longer EMAs, and discarded repetitive EMAs as the novelty wore off. By design, there is no notification when an EMA is fired. Participants need to actively check their phone to answer scheduled EMA questions. The EMA compliance data (see Fig. 1b) shows that there are no significant differences between primary and secondary phone users. It indicates that secondary phone users also used the study phone when they were taking the phone with them. Therefore, the study phone can capture the participants' daily behavior even it was not their primary phone.

In summary, Fig. 1 shows the cost of collecting continuous and EMA data across a 10-week study. There is a small difference between primary and secondary collectors for continuous sensing and EMA data, but the compliance reported above is promising and gives confidence in the analysis discussed in section "Results".

StudentLife App and Sensing System

In what follows, we describe the design of the StudentLife app and sensing system, as shown in Fig. 2.

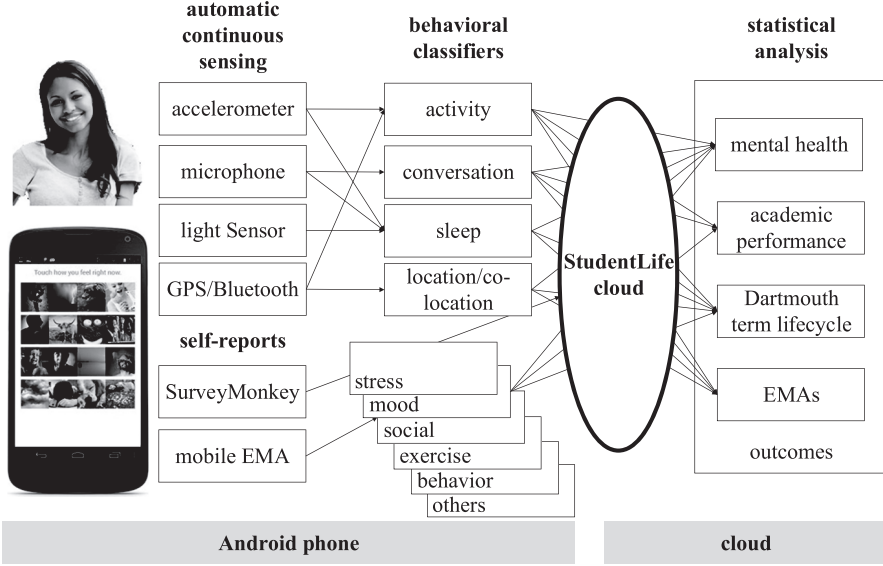


Fig. 2 StudentLife app, sensing and analytics system architecture

Automatic and Continuous Sensing

We build on our prior work on the BeWell App [33] to provide a framework for automatic sensing in StudentLife. The StudentLife app automatically infers activity (stationary, walking, running, driving, cycling), sleep duration, and sociability (i.e., the number of independent conversations and their durations). The app also collects accelerometer, proximity, audio, light sensor readings, location, colocation, and application usage. The inferences and other sensor data are temporarily stored on the phone and are efficiently uploaded to the StudentLife cloud when users recharge their phones under WiFi. In what follows, we discuss the physical activity, sociability/conversation and sleep inferences computed on the phone which represent important health well-being indicators [33].

Activity Detection We use the physical activity classifier from our prior work [33, 35] to infer stationary, walking, running, driving and cycling based on features extracted from accelerometer streams. The activity classifier extracts features from the preprocessed accelerometer stream, then applies a decision tree to infer the activity using the features. The activity classifier achieves overall 94% of accuracy [35]. (Note, we conducted our study before Google announced the availability of an activity recognition service for Android phones). We extend our prior work to compute a daily activity duration, and indoor and outdoor mobility measures, discussed as follows. The activity classifier generates an activity label every 2 s. We are only interested in determining whether a participant is moving. For each

10-min period, we calculate the ratio of non-stationary inferences. If the ratio is greater than a threshold, we consider this period active, meaning that the user is moving. We add up all the 10-min active periods as the daily activity duration. Typically, students leave their dorms in the morning to go to various buildings on campus during the day. Students spend a considerable amount of time in buildings (e.g., cafes, lecture rooms, gym). We consider the overall mobility of a student consists of indoor and outdoor mobility. We compute the outdoor mobility (*aka* traveled distance) as the distance a student travels around campus during the day using periodic GPS samples. Indoor mobility is computed as the distance a student travels inside buildings during the day using WiFi scan logs. Dartmouth College has WiFi coverage across all campus buildings. As part of the study, we collect the locations of all APs in the network, and the Wi-Fi scan logs including all encountered BSSIDs, SSIDs, and their signal strength values. We use the BSSIDs and signal strength to determine if a student is in a specific building. If so, we use the output of activity classifier's walk inference to compute the activity duration as a measure of indoor mobility. Note, that Dartmouth's network operations provided access to a complete AP map of the campus wireless network as part of the IRB.

Conversation Detection StudentLife implements two classifiers on the phone for audio and speech/conversation detection: an audio classifier to infer human voice, and a conversation classifier to detect conversation. We process audio on the fly to extract and record features. We use the privacy-sensitive audio and conversation classifiers developed in our prior work [33, 41]. Note, the audio classification pipeline never records conversation nor analyses content. We first segment the audio stream into 15-ms frames. The audio classifier then extracts audio features, and uses a two-state hidden Markov model (HMM) to infer speech segments. Our classifier does not implement speaker identification. It simply infers that the user is “around conversation” using the output of the audio classifier as an input to a conversation classifier. The output of the classification pipeline captures the number of independent conversations and their duration. We consider the frequency and duration of conversations around a participant as a measure of sociability. Because not all conversations are social, such as lectures and x-hours (i.e., class meetings outside lectures), we extend our conversation pipeline in the cloud to remove conversations associated with lectures and x-hours. We use student location to determine if they attend lectures and automatically remove the conversation data correspondingly from the dataset discussed in section “StudentLife Dataset”. We also keep track of class attendance for all students across all classes, as discussed in section “Results”.

Sleep Detection We implement a sleep classifier based on our previous work [16, 33]. The phone unobtrusively infers sleep duration without any special interaction with the phone (e.g., the user does not have to sleep with the device). The StudentLife sleep classifier extracts four types of features: light features, phone usage features including the phone lock state, activity features (e.g., stationary), and sound features from the microphone. Any of these features alone is a weak classifier for sleep duration because of the wide variety of phone usage patterns.

Our sleep model combines these features to form a more accurate sleep model and predictor. Specifically, the sleep model assumes that sleep duration (Sl) is a linear combination of these four factors: $Sl = \sum_{i=1}^4 \alpha_i \cdot F_i$, $\alpha_i \geq 0$ where α_i is the weight of the corresponding factor. We train the model using the method described in [16] with an accuracy of ± 32 min to the ground truth. We extend this method to identify the sleep onset time by looking at when the user is sedentary in term of activity, audio, and phone usage. We compare the inferred sleep onset time from a group of 10 students who use the Jawbone UP during the study to collect sleep data. Our method predicts bedtime where 95% of the inferences have an accuracy of ± 25 min of the ground truth. The output of our extended sleep classifier is the onset of sleep (i.e., bedtime), sleep duration and wake up time.

MobileEMA

We use in-situ ecological momentary assessment (EMA) [45] to capture additional human behavior beyond what the surveys and automatic sensing provide. The user is prompted by a short survey (e.g., the single item [47] stress survey as shown in Fig. 3b) scheduled at some point during their day. We integrate an EMA component

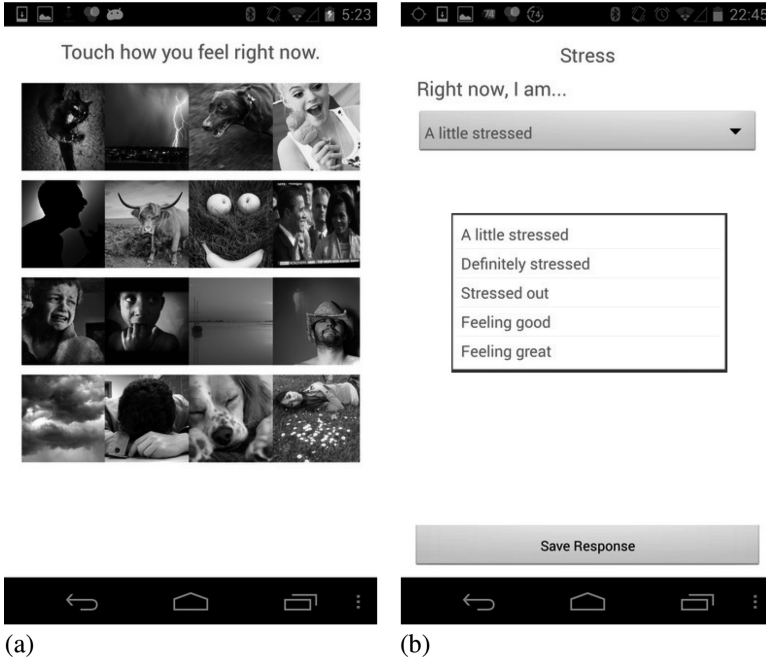


Fig. 3 MobileEMA: first the PAM popup fires followed by one of the StudentLife EMAs—in this example the single item stress EMA. (a) PAM EMA. (b) Stress EMA

into the StudentLife app based on extensions to Google PACO [5]. PACO is an extensible framework for quantified self experiments based on EMA. We extend PACO to incorporate:

- *photographic affect meter (PAM)* [39] to capture participant's instantaneous mood;
- *pop-up EMAs* to automatically present a short survey to the user when they unlock or use the phone; and,
- *EMA schedule and sync* feature to automatically push a new EMA schedule to all participants and synchronize the new schedule with StudentLife cloud.

PACO is a self-contained and complex backend app and service. We extend and remove features and integrate the EMA component into the StudentLife app and cloud. We set up EMA questions and schedules using the PACO server-side code [5]. The cloud pushes new EMA questions to the phones. The StudentLife app sets up an alarm for each EMA in the list and fires it by pushing it to the users' phone screen as a pop-up. We implement PAM [39] on the Nexus 4 as part of the EMA component. PAM presents the user with a randomized grid of 16 pictures from a library of 48 photos. The user selects the picture that best fits their mood. Figure 3a shows the PAM pop-up asking the user to select one of the presented pictures. PAM measures affect using a simple visual interface. PAM is well suited to mobile usage because users can quickly click on a picture and move on. Each picture represents a 1–16 score, mapping to the Positive and Negative Affect Schedule (PANAS) [55]. PAM is strongly correlated with PANAS ($r = 0.71, p < 0.001$) for positive affect. StudentLife schedules multiple EMAs per day. We took the novel approach of firing PAM before showing one of the scheduled EMAs (e.g., stress survey). Figure 3b shows an EMA test after the PAM pop-up. We are interested in how students' mood changes during the day. By always preceding any EMA with PAM, we guarantee a large amount of affect data during the term.

StudentLife Dataset

Using the StudentLife system described in section “StudentLife Sensing System Section”, we collect a dataset from all subjects including automatic sensor data, behavioral interferences, and self-reported EMA data. Our ground truth data includes behavioral and mental health outcomes computed from survey instruments detailed in Table 1, and academic performance from spring term and cumulative GPA scores provided by the registrar. We discuss three *epochs* that are evident in the StudentLife dataset. We use these epochs (i.e., *day* 9 am–6 pm, *evening* 6 pm–12 am, *night* 12 am–9 am) as a means to analyze some of the data, as discussed in section “Results”. The StudentLife dataset is publicly available [6].

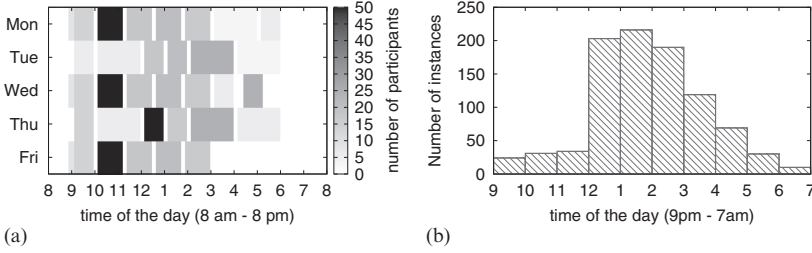


Fig. 4 Statistics on class meeting times and sleep onset time (i.e., bedtime). **(a)** Meeting time for all classes over the term. **(b)** Sleep onset time distribution for all students over the term

Automatic Sensing Data We collect a total of 52.6 GB of sensing inference data from smartphones over 10 weeks. The data consist of: (1) activity data, including activity duration (total time duration the user moves per day), indoor mobility and the total traveled distance (i.e., outdoor mobility) per day; (2) conversation data, including conversation duration and frequency per day; (3) sleep data, including sleep duration, sleep onset and waking time; and finally (4) location data, including GPS, inferred buildings when the participant is indoors, and the number of co-located Bluetooth devices.

Epochs Students engage in different activities during the day and night. As one would expect, sleep and taking classes dominate a student’s week. Figure 4a shows the collective timetable of class meetings for all the classes taken by the students in the study. The darker the slot, the greater proportion of students taking classes in the slot. We can observe that Monday, Wednesday, Friday slots from 10:00–11:05 am and the x-period on Thursday 12:00–12:50 pm are dominant across the week; this is the teaching time for the CS65 Smartphone Programming class which all students in the study are enrolled in. Figure 4a clearly indicates that the timetable of all classes ranges from 9 am to 6 pm—we label this as the *day epoch*. Students are not taking classes for the complete period. Many class, social, sports, and other activities take place during the day epoch but class is dominant. The next dominant activity is sleep. Students go to bed at different times. Figure 4b shows the distribution of bedtime for all students across the term. We see that most students go to bed between 12 am and 4 am but the switch from evening to night starts at 12 am, as shown in Fig. 4b. We label the period between 12 am and 9 am as the *night epoch*, when most students are working, socializing or sleeping—but sleep is the dominant activity. We consider the remaining period between the end of classes (6 pm) and sleep (12 am) as the *evening epoch*. We hypothesize that this is the main study and socialization period during weekdays. We define these three epochs as a means to analyze data, as discussed in section “Results”. We appreciate that weekdays are different from weekends but consider epochs uniformly across the complete week. We also look for correlations in complete days (e.g., Monday) and across epochs (i.e., Monday day, evening and night).

EMA Data Students respond to psychological and behavioral EMAs on their smartphones that are scheduled, managed, and synchronized using the MobileEMA component integrated into StudentLife app. We collect a total of 35,295 EMA and PAM responses from 48 students over 10 weeks. EMA and PAM data are automatically uploaded to the cloud when students recharge their phones under WiFi. Students respond to a number of scheduled EMAs including stress (stress EMA), mood (mood EMA), sleep duration (sleep EMA)(which we use to confirm the performance of our sleep classifier), the number of people students encountered per day (social EMA), physical exercise (exercise EMA), time spent on different activities (activity EMA), and short personality item (behavior EMA). All EMAs were either existing validated EMAs (e.g., single item stress measure [47]) found in the literature, or provided by psychologist on the team (e.g., mood EMA).

Survey Instrument Data Table 1 shows the set of surveys for measuring behavioral and mental well-being and personality traits we administer as part of our pre-post stages, as discussed in section “Study Design”. These questionnaires provide an assessment of students’ depression, perceived stress, flourishing (i.e., self-perceived success), loneliness, and personality. Students complete surveys using SurveyMonkey [7] 1 day prior to study commencement, and complete them again 1 day after the study. Surveys are administered on the phone and stored in the StudentLife cloud (Fig. 2). In what follows, we overview each instrument. The Patient Health Questionnaire (PHQ-9) [32] is a depression module that scores each of the 9 DSM-IV criteria as 0 (not at all) to 3 (nearly every day). It is validated for use in primary care. Table 2 shows the interpretation of the scale and the number of students that fall into each category for pre-post assessment. The perceived stress scale (PSS) [19] measures the degree to which situations in a person’s life are stressful. Psychological stress is the extent to which a person perceives the demands on them exceed their ability to cope [19]. Perceived stress is scored between 0 (least stressed) to 40 (most stressed). The perceived stress scale can only be used for comparisons within a sample—in our case 48 students. The flourishing scale [21] is an 8-item summary measure of a person’s self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism. The scale provides a single psychological well-being score. Flourishing is scored between 8 (lowest) to 56 (highest). A high score represents a person with many psychological resources and strengths. The UCLA loneliness (version 3) [44] scale scores between 20 (least lonely) to 80 (most lonely). The loneliness scale is a 20-item scale designed to

Table 2 PHQ-9 depression scale interpretation and pre-post class outcomes

Depression severity	Minimal	Minor	Moderate	Moderately severe	Severe
Score	1–4	5–9	10–14	15–19	20–27
Number of students (pre-survey)	17	15	6	1	1
Number of students (post-survey)	19	12	3	2	2

Table 3 Statistics of mental well-being surveys

Survey outcomes	Pre-study			Post-study		
	Participants	Mean	Std	Participants	Mean	Std
Depression	40	5.8	4.9	38	6.3	5.8
Flourishing	40	42.6	7.9	37	42.8	8.9
Stress	41	18.4	6.8	39	18.9	7.1
Loneliness	40	40.5	10.9	37	40.9	10.5

measure a person’s subjective feelings of loneliness as well as feelings of social isolation. Low scores are considered a normal experience of loneliness. Higher scores indicate a person is experiencing severe loneliness. Table 3 shows the pre-post measures (i.e., mean and standard deviation) for each scored survey for all students. We discuss these assessments in section “Results”.

Academic Data We have access to transcripts from the registrar’s office for all participants as a means to evaluate their academic performance. We use spring and cumulative GPA scores as ground truth outcomes. Undergraduates can receive an A–E grade or I (incomplete). Students who get an Incomplete must agree to complete the course by a specific date. GPA ranges from 0 to 4. For the CS65 smartphone programming class we had all the assignment and project deadlines—no midterms or finals are given in this class. Students provide deadlines of their other classes at the exit interview from their calendars or returned assignments or exams.

Results

In what follows, we discuss the main results from the StudentLife study. We identify a number of significant correlations between objective sensor data from smartphones and mental well-being and academic performance outcomes. We also identify a Dartmouth term lifecycle that captures the impact of the term on behavioral measures representing an aggregate term signature experienced by all students.

Correlation with Mental Health

We first consider correlations between automatic and objective sensing data from smartphones and mental well-being. We also discuss results from correlations between EMA data. Specifically, we report on a number of significant correlations between sensor and EMA data and pre-post survey ground truth outcomes for

Table 4 Correlations between automatic sensor data and PHQ-9 depression scale

Automatic sensing data	<i>r</i>	<i>p</i> -value
Sleep duration (pre)	−0.360	0.025
Sleep duration (post)	−0.382	0.020
Conversation frequency during day (pre)	−0.403	0.010
Conversation frequency during day (post)	−0.387	0.016
Conversation frequency during evening (post)	−0.345	0.034
Conversation duration during day (post)	−0.328	0.044
Number of co-locations (post)	−0.362	0.025

depression (PHQ-9), flourishing, perceived stress, and loneliness scales, as discussed in section “StudentLife Dataset” and shown in Table 3. We calculate the degree of correlation between sensing/EMA data and outcomes using the Pearson correlation [18] where r ($-1 \leq r \leq 1$) indicates the strength and direction of the correlation, and p the significance of the finding.

PHQ-9 Depression Scale Table 2 shows the pre-post PHQ-9 depression severity for the group of students in the study. The majority of students experience minimal or minor depression for pre-post measures. However, 6 students experience moderate depression and 2 students are moderately severe or severely depressed at the start of term. At the end of term more students (4) experience either moderately severe or severely depressed symptoms. We find a number of significant correlations ($p \leq 0.05$) between sleep duration, conversation frequency and duration, colocation (i.e., number of Bluetooth encounters) and PHQ-9 depression, as shown Table 4. An inability to sleep is one of the key signs of clinical depression [3]. We find a significant negative correlation between sleep duration and pre ($r = -0.360, p = 0.025$) and post ($r = -0.382, p = 0.020$) depression; that is, students that sleep less are more likely to be depressed. There is a known link between lack of sleep and depression. One of the common signs of depression is insomnia or an inability to sleep [3]. Our findings are inline with these studies on depression [3]. However, we are the first to use automatic sensor data from smartphones to confirm these findings. We also find a significant negative association between conversation frequency during the day epoch and pre ($r = -0.403, p = 0.010$) and post ($r = -0.387, p = 0.016$) depression. This also holds for the evening epoch where we find a strong relationship ($r = -0.345, p = 0.034$) between conversation frequency and depression score. These results indicate that students that have fewer conversational interactions are more likely to be depressed. For conversation duration, we find a negative association ($r = -0.328, p = 0.044$) during the day epoch with depression. This suggests students who interact less during the day period when they are typically social and studying are more likely to experience depressive symptoms. In addition, students that have fewer co-locations with other people are more likely ($r = -0.362, p = 0.025$) to have a higher PHQ-9 score. Finally, we find a significant positive correlation ($r = 0.412, p = 0.010$) between

Table 5 Correlations between automatic sensor data and flourishing scale

Automatic sensing data	<i>r</i>	<i>p</i> -value
Conversation duration (pre)	0.294	0.066
Conversation duration during evening (pre)	0.362	0.022
Number of co-locations (post)	0.324	0.050

Table 6 Correlations between automatic sensor data and perceived stress scale (PSS)

Automatic sensing data	<i>r</i>	<i>p</i> -value
Conversation duration (post)	−0.357	0.026
Conversation frequency (post)	−0.394	0.013
Conversation duration during day (post)	−0.401	0.011
Conversation frequency during day (pre)	−0.524	0.001
Conversation frequency during evening (pre)	−0.386	0.015
Sleep duration (pre)	−0.355	0.024

the validated single item stress EMA [47] and the post PHQ-9 scale. This indicates that people that are stressed are also more likely to experience depressive symptoms, as shown in Table 7.

Flourishing Scale There are no literal interpretation of flourishing scale, perceived stress scale (PSS) and UCLA loneliness scale instruments, as discussed in section “StudentLife Dataset”. Simply put, however, the higher the score the more flourishing, stressed and lonely a person is. We find a small set of correlations (see Table 5) between sensor data and flourishing. Conversation duration has a weak positive association ($r = 0.294, p = 0.066$) during the 24 h day with flourishing. With regard to conversation during the evening epoch we find a significant positive association ($r = 0.362, p = 0.022$) with flourishing. We also find that students with more co-locations ($r = 0.324, p = 0.050$) are more flourishing. These results suggest that students that are more social and around people are more flourishing. Finally, positive affect computed from the PAM self-report has significant positive correlation ($r = 0.470, p = 0.002$) with flourishing, as shown in Table 7. This is as we would imagine. People who have good positive affect flourish.

Perceived Stress Scale Table 6 shows the correlations between sensor data and perceived stress scale (PSS). Conversation frequency ($r = -0.394, p = 0.013$) and duration ($r = -0.357, p = 0.026$) show significantly negative correlation with post perceived stress. In addition, we see more significant negative associations if we just look at the day epoch. Here, conversation frequency ($r = -0.524, p = 0.001$) and duration ($r = -0.401, p = 0.011$) exhibit significant and strong negative correlations with pre and post measure of perceived stress, respectively. This suggests students in the proximity of more frequent and longer conversations during the day epoch are less likely to feel stressed. We cannot distinguish between social and work study conversation, however. We hypothesize that students work collaborative in study groups. And these students make more progress and are less stressed. There is also strong evidence that students that are around more

Table 7 Correlations between EMA data and mental well-being outcomes

Mental health outcomes	EMA	<i>r</i>	<i>p</i> -value
Flourishing scale (pre)	Positive affect	0.470	0.002
Loneliness (post)	Positive affect	−0.390	0.020
Loneliness (post)	Stress	0.344	0.037
PHQ-9 (post)	Stress	0.412	0.010
Perceived stress scale (pre)	Positive affect	−0.387	0.012
Perceived stress scale (post)	Positive affect	−0.373	0.019
Perceived stress scale (pre)	Stress	0.458	0.003
Perceived stress scale (post)	Stress	0.412	0.009

Table 8 Correlations between automatic sensor data and loneliness scale

Automatic sensing data	<i>r</i>	<i>p</i> -value
Activity duration (post)	−0.388	0.018
Activity duration for day (post)	−0.326	0.049
Activity duration for evening (post)	−0.464	0.004
Traveled distance (post)	−0.338	0.044
Traveled distance for day (post)	−0.336	0.042
Indoor mobility for day (post)	−0.332	0.045

conversations in the evening epoch are less stressed too. Specifically, there is strong negative relationship ($r = -0.386, p = 0.015$) between conversation frequency in the evening epoch and stress. There is also a link between sleep duration and stress. Our results show that there is a strong negative association ($r = -0.355, p = 0.024$) between sleep duration and perceived stress. Students that are getting more sleep experience less stress. Finally, we find significant positive ($r = 0.458, p = 0.003$) and negative correlations ($r = -0.387, p = 0.012$) between self-reported stress levels and positive affect (i.e., PAM), respectively, and the perceived stress scale. There is a strong connection between daily reports of stress over the term and the pre-post perceived stress scale, as shown in Table 7. Similarly, students that report higher positive affect tend to be less stressed.

Loneliness Scale We find a number of links between activity duration, distance travelled, indoor mobility and the loneliness scale, as shown in Table 8. All our results relate to correlations with post measures. Activity duration during a 24 h day has a significant negative association ($r = -0.388, p = 0.018$) with loneliness. We can look at the day and evening epochs and find correlations. There is a negative correlation ($r = -0.464, p = 0.004$) between activity duration in the evening epoch and loneliness. Distance traveled during the complete day ($r = -0.338, p = 0.044$) and the day epoch ($r = -0.336, p = 0.042$) show trends with being lonely. Indoor mobility during the day epoch has strong negative links ($r = -0.332, p = 0.045$) to loneliness. Indoor mobility is a measure of how much a student is moving in buildings during the day epoch. Students that are less active and therefore less mobile are more likely to be lonely. It is difficult to speculate about cause and

effect. Maybe these students move around less are more isolated (e.g., stay in their dorm) because they have less opportunity to meet other students outside of class. These students could feel lonely and therefore more resigned not to seek out the company of others. There is also no evidence that people who interact with others regularly do not experience loneliness. This supports our lack of findings between conversation and loneliness. The PAM EMA data (positive affect) has a strong negative association ($r = -0.390, p = 0.020$) with positive affect. In addition, stress self-reports positively correlate ($r = 0.344, p = 0.037$) with loneliness. Students who report higher positive affect and less stress tend to report less loneliness, as shown in Table 7.

Predicting Academic Performance

We use a subset of the StudentLife dataset to analyze and predict academic performance. We only use undergraduate students' ($N = 30$) data because only undergraduates have GPAs. In contrast, Dartmouth graduate students do not have GPAs and only receive High Pass, Pass, Low Pass or No Credit in their classes. We propose new methods to automatically infer *study* (i.e., study duration and focus) and *social* (i.e., partying) *behaviors* using passive sensing from smartphones [52]. We use time series analysis of behavioral states to predict cumulative GPA. We use linear regression with lasso regularization to identify non-redundant predictors among a large number of input features and use these features to predict students' cumulative GPA.

Assessing Study and Social Behavior The StudentLife dataset provides a number of low-level behaviors (e.g., physical activity, sleep duration, and sociability based on face-to-face conversational data) but offers no higher level data related to study and social behaviors, which are likely to impact academic performance. We attribute meanings or semantics to locations—called behavioral spaces [52] as a basis to better understand study and social behaviors. That is, we extract high level behaviors, such as studying (e.g., study duration and focus) and social (e.g., partying) behaviors by fusing multiple sensor streams with behavioral spaces.

We use behavioral space information to determine study behavior [52]. Each student takes three classes, which are scheduled at specific periods during the week [2]. Students' transcripts indicate what classes they took. The registrar office has the schedule and location for each class. We use location, date (i.e., weekday M-F) and time to automatically determine if a student attends a class or not, checking the dwell time at the location at least equals 90% of the scheduled period (e.g., 110 min). Using this approach the phone can automatically determine the classes a student is taking and their attendance rates.

We heuristically determine if a student's dwell time at a study areas (e.g., library, labs, study rooms, cafes where student primarily work) is at least 20 min. We consider periods shorter than 20 min are less likely to be real study periods.

In addition to dwell time, we use activity and audio attributes to determine a student's level of focus at a study area. The value of activity indicates how often the phone moves—the person is either moving around in the study area or stationary but using the phone. We consider a number of scenarios. If a student is in a study (e.g., a library) and moves around we consider this contributes to a lack of focus. If the phone is mostly stationary in a study area, we consider this contributes to focus. We also use the audio attribute to determine the level of ambient noise in study areas. We consider quiet environments may contribute to study focus and noisy environments do not. In term of focus, a higher activity value indicates that the student moves around less and thus is more focused and a higher audio value indicates that the student is in a quieter environment which is more conducive to being focused. We do not combine these values but use them as independent variables in the analysis section.

We consider behavioral spaces (e.g., Greek houses, dorms) and their attributes to infer if a student is partying [52]. If a student is in a party we assume that they will be moving and around acoustic sound of conversation or music. We also consider the day of the week as being significant for the fraternity and sorority parties (i.e., Wednesday, Friday and Saturday). We discard dwell times under 30 min at partying locations.

We partition each Greek house dwell periods (i.e., visit or stay) into 10-min windows and calculate audio and activity attributes. We hypothesize that the audio and the activity attributes should be significantly different when the student is partying or not partying. We use k-means clustering [56] to find the partying thresholds for both the audio (e.g., music or being surrounded by a large group of people) and activity (e.g., dancing) attributes.

Capturing Behavioral Change We extract behavioral change features from the low-level automatic sensing (e.g., sleep duration) and EMA data (e.g., stress) and high-level study and social behaviors discussed in the previous section. We create time series of each behavior for each student. The behavior time series samples each behavior each day. Each time series summarizes a different behavior (e.g., physical activity, conversation frequency and duration, sleep, social behavior, and study behaviors). In order to understand behavior changes across the term, we propose two features [52]: *behavioral slope*, which captures the magnitude of change (e.g., increase or decrease in sleep) over the complete term as well as the first and second half of the term for all students—from the start of term to the midterm point, and then from the midterm point to the end of term; and *behavioral breakpoints*, which capture the specific points in the term where individual behavior change occurs—the number of breakpoints a student experiences indicates the rate of change that occurs. The method to extract these behavioral change features are described in detail in [52].

Predicting Cumulative GPA Predicting GPA is a regression problem; that is, predicting an outcome variable (i.e., GPA) from a set of input predictors (i.e., features). We evaluate various regression models such as regularized linear regression, regression trees, and support vector regression using cross-validation. We

select the *Lasso* (Least Absolute Shrinkage and Selection Operator) [48] regularized linear regression model as our predictive model. Lasso is a method used in linear regression; that is, Lasso minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Considering we have a large number of features, collinearity needs to be addressed. There are two categories of methods that address collinearity: feature selection and feature transformation. Lasso regularization is one of the feature selection methods. *Lasso* solves the following optimization problem:

$$\min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

where N is the number of observations; y_i is the ground truth of observation i ; x_i is the p degree feature vector at observation i ; λ is a nonnegative regularization parameter, which controls the number of nonzero components of β (i.e., number of the selected features); β_0 is the intercept; and β is the weight vector. The regularization parameter λ is selected using cross-validation. The optimization problem is essentially to minimize the mean square error $\frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2$ of fitting while keeping the model as simple as possible (i.e., select a minimal number of features to avoid overfitting). Thus, *Lasso* automatically selects more relevant features (i.e., predictors) and discards redundant features to avoid overfitting.

We use the mean absolute errors (MAE), the coefficient of determination (R^2) [15], and Pearson correlation to measure the performance of outcome prediction. MAE measures how close predictions are to the outcomes. The mean absolute error is given by $\text{MAE} = \frac{1}{n} \sum_{i=1}^N |y_i - \beta_0 - x_i^T \beta|$. Smaller MAE is preferred because it indicates that the predictions are closer to the ground truth. R^2 is another statistic that measures the goodness of fit of a model and indicates how much of the variance our model explains. R^2 ranges from 0 to 1, where 1 indicates that the model perfectly fits the data. R^2 can be seen to be related to the unexplained variance where $R^2 = 0$ if the feature vector X tells us nothing about the outcome. We use Pearson correlation to measure the linear relations between the ground truth and the predictive outcome.

We apply leave-one-subject-out cross validation [30] to determine the parameters for *Lasso* and the weights for each feature. In order to make the weight regularization work properly, each feature is scaled within the range $[0, 1]$. Selected features have non-zero weights. The MAE of our predicted cumulative GPA is 0.179, indicating that the predictions are within ± 0.179 of the groundtruth. The R^2 is 0.559, which indicates that the features can explain 55.9% of the GPA variance. The predicted GPA strongly correlates with the ground truth with $r = 0.81$ and $p < 0.001$, which further indicates that our predictions can capture outcome differences using the given features.

Table 9 shows the selected features to predict the cumulative GPAs and their weights. Interestingly, *lasso* selects a single long term measure (i.e., conscientious personality trait), two self-report time series features (i.e., affect and stress), and three automatic sensing data behaviors (i.e., conversational and study behavior).

Table 9 Lasso selected GPA predictors and weights

	Features	Weight
Sensing	Conversation duration night breakpoint	0.3467
	Conversation duration evening term-slope	−0.6100
	Study duration	0.0728
EMA	Positive affect	0.0930
	Positive affect post-slope	−0.1215
	Stress term-slope	−2.6832
Survey	Conscientiousness	0.0449

The weights indicate the strength of the predictors. Students who have better GPAs are more conscientious, study more, experience positive moods (e.g., joy, interest, alertness) across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period between 6–12 pm, and experience later change (i.e., a behavioral breakpoint) in their conversation duration pattern.

Dartmouth Term Lifecycle

We analyze the Dartmouth term lifecycle using both sensing data and self-reported EMA data. Figure 5a–c shows key behavioral measures and activities over the complete term. Figure 5a shows EMA data for stress and positive affect (PA), and automatic sensing data for sleep duration. Figure 5b shows continuous sensing trends specifically activity duration, and conversation duration and frequency. Finally, Fig.5c shows location based data from GPS and WiFi, specifically, attendance across all classes, the amount of time students spent in their dorms or at home, and visits to the gym. We hypothesize that these sensing, EMA and location based curves collectively represent a “Dartmouth term lifecycle”. Whether these trends could be observed across a different set of students at Dartmouth or more interestingly at a different institution is future work. In what follow we discuss workload across the term, mental well-being using EMA data (i.e., stress and positive affect) and automatic sensing data measures.

Academic Workload We use the number of assignment deadlines as a measure of the academic workload of students. We collect class deadlines during exit interviews and validate them against students’ calendars and returned assignments dates. Figure 5 shows the average number of deadlines for all student across each week of the term. The number of deadlines peaks during the mid-term period in weeks 4 and 5. Interestingly, many classes taken by the students do not have assignment deadlines during week 8. Final projects and assignments are due in the last week of term before finals, as shown in Fig. 5a. As discussed before, all study participants take the same CS65 Smartphone Programming class, for which they share the same

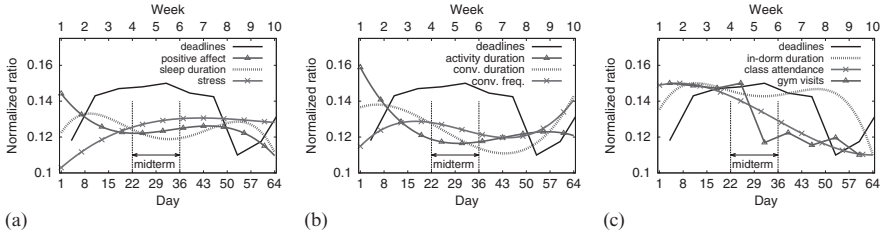


Fig. 5 Dartmouth term lifecycle: collective behavioral trends for all students over the term. **(a)** EMA and sleep data. **(b)** Automatic sensing data. **(c)** Location-based data

deadlines. Among all CS65’s lab assignment, Lab 4 is considered to be the most challenging programming assignment. In the last week of term the students need to give final presentations and live demos of group projects for the smartphone programming class. The students are told that app developed for the demo day has to work to be graded. The demo is worth 30% of their overall grade.

Self Reported Stress and Mood Figure 5a shows the average daily stress level and positive affect over the term for all subjects as polynomial curves. Students are more stressed during the mid-term (days 22–36) and finals periods. The positive affect results show a similar trend. Students start the term with high positive affect, which then gradually drops as the term progresses. During the last week of term, students may be stressed because of finals and class projects, with positive affect dropping to its lowest point in the term. Overall, the results indicate that the 10-week term is stressful for students as workload increases. Figure 5a clearly shows that students return to Dartmouth after spring break feeling the most positive about themselves, the least stressed, the most social in terms of conversation duration and the most active (as shown in Fig. 5b). As the term progresses toward mid-term week, positive affect and activity duration plunge and remain at low levels until the final weeks where positive affect drops to its lowest point.

Automatic Sensing Data We also study behavioral patterns over the term by analyzing automatic sensing data. We plot the polynomial fitting curves for sleep duration, activity duration, conversation duration, conversation frequency, as shown Fig. 5b, and location visiting patterns in Fig. 5c. Our key findings are as follows. We observe from Fig. 5a that sleep peaks at the end of the first week and then drops off and is at its lowest during the mid-term exam weeks. Sleep then improves until the last week of term when it plummets to its lowest point in the cycle. As shown in Fig. 5b students start the term with larger activity duration, which gradually drops as they become busier with course work and other term activities. Finally, the activity duration increases a little toward the end of term. Activity duration reaches its lowest point on day 36 when students are focused on completing the Lab 4 assignment—considered the most demanding assignment in the smartphone programming class.

The student’s level of face-to-face sociability starts high at the start of term, then we observe an interesting conversation pattern, as shown in Fig. 5b. As the

term intensifies, conversation duration drops almost linearly until week 8, and then rebounds to its highest point at the end of term. Conversely, the frequency of conversation increases from the start of term until the start of midterms, and then it drops and recovers toward the end of term. We speculate that sociability changes from long social/study related interactions at the start of term to more business-like interactions during midterms when students have shorter conversations. At the end of term, students are having more frequent, longer conversations.

Figure 5c provides a number of interesting insights based on location based data. As the term progresses and deadlines mount the time students spend at the significant places in their lives radically changes. Visits to the gym plummet during midterm and never rebound. The time students spend in their dorm is low at the start of term perhaps due to socializing then remains stable but drops during midterm. At week 8 time spent in dorms drops off and remains low until the end of term. The most interesting curve is class attendance. We use location data to determine if students attend classes. We consider 100% attendance when all students attend all classes and x-hours (if they exist). The term starts with 75% attendances and starts dropping at week 3. It steadily declines to a point at the end of term where only 25% of the class are attending all their classes. Interestingly, we find no correlation between class attendance and academic performance. We speculate that students increasingly start missing classes as the term progresses and the work load rises. However, absence does not positively or negatively impact their grades. We put this down to their self learning ability but plan to study this further as part of future work.

It is difficult in this study to be concrete about the cause and effect of this lifecycle. For example, stress or positive affect could have nothing to do with workload and everything to do with hardship of some sort (e.g., campus adjustment, roommate conflicts, health issues). We speculate the intensive workload compressed into a 10 week term puts considerable demands on students. Those that excel academically develop skills to effectively manage workload, social life and stress levels.

Conclusion

In this paper, we presented the StudentLife sensing system and results from a 10-week deployment. We discuss a number of insights into behavioral trends, and importantly, correlations between objective sensor data from smartphones and mental well-being and predicting undergraduate students' cumulative GPA for a set of students at Dartmouth College. To the best of our knowledge, this is the first-of-its-kind smartphone sensing system and study. Providing feedback of hidden states to students and other stakeholders might be beneficial, but there are many privacy issues to resolve. Students, deans, and clinicians on campus all care about the health and well-being of the student body. In this study, the professor running the study had access to survey outcomes, sensing data, and EMAs for students. In two cases,

the professor intervened and did not give failing grades to students who failed to complete a number of assignments and missed lectures for several weeks. Rather, they were given incomplete grades and completed assignments over the summer. However, in other classes these students took, their professors did not have such data available and these students received failing grades. While access to such data is under IRB and cannot be shared, the data and intervention in grading enabled those students to return to campus the following fall. If they had received 3 failing grades, they would have been suspended for one term.

References

1. CS65 Smartphone Programming (2013). <http://www.cs.dartmouth.edu/~campbell/cs65/cs65.html>
2. Dartmouth College Weekly Schedule Diagram (2013). <http://oracle-www.dartmouth.edu/dartgroucho/timetabl.diagram>
3. Depression (2016). <http://www.nlm.nih.gov/health/topics/depression/index.shtml>
4. funf-open-sensing-framework (2013). <https://code.google.com/p/funf-open-sensing-framework/>
5. PACO (2013). <https://code.google.com/p/paco/>
6. StudentLife Dataset (2014). <http://studentlife.cs.dartmouth.edu/>
7. SurveyMonkey (2013). <https://www.surveymonkey.com/>
8. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* **7**(6), 643–659 (2011)
9. Aldwin, C.M.: Stress, coping, and development: An integrative perspective. Guilford Press (2007)
10. Bang, S., Kim, M., Song, S.K., Park, S.J.: Toward real time detection of the basic living activity in home using a wearable sensor and smart home sensors. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 5200–5203. IEEE (2008)
11. Bardram, J.E., Frost, M., Szántó, K., Marcu, G.: The monarca self-assessment system: a persuasive personal monitoring system for bipolar patients. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 21–30. ACM (2012)
12. Bergman, R.J., Bassett Jr, D.R., Klein, D.A.: Validity of 2 devices for measuring steps taken by older adults in assisted-living facilities. *Journal of physical activity & health* **5** (2008)
13. Bravata, D.M., Smith-Spangler, C., Sundaram, V., Gienger, A.L., Lin, N., Lewis, R., Stave, C.D., Olkin, I., Sirard, J.R.: Using pedometers to increase physical activity and improve health: a systematic review. *Jama* **298**(19), 2296–2304 (2007)
14. Burns, M.N., Begale, M., Duffecy, J., Gergle, D., Karr, C.J., Giangrande, E., Mohr, D.C.: Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research* **13**(3) (2011)
15. Cameron, A.C., Windmeijer, F.A.: R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics* **14**(2), 209–220 (1996)
16. Chen, Z., Lin, M., Chen, F., Lane, N.D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., Campbell, A.T.: Unobtrusive sleep monitoring using smartphones. In: Proc. of PervasiveHealth (2013)
17. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., et al.: The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, IEEE* **7**(2), 32–41 (2008)

18. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge (1988)
19. Cohen, S., Kamarck, T., Mermelstein, R.: A global measure of perceived stress. *Journal of health and social behavior* pp. 385–396 (1983)
20. Cowie, R., Douglas-Cowie, E.: Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1989–1992. IEEE (1996)
21. Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.w., Oishi, S., Biswas-Diener, R.: New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research* **97**(2), 143–156 (2010)
22. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. *Personal and ubiquitous computing* **10**(4), 255–268 (2006)
23. Fausett, L., Elwasif, W.: Predicting performance from test scores using backpropagation and counterpropagation. In: *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 5, pp. 3398–3402 vol.5 (1994). doi:10.1109/ICNN.1994.374782
24. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, D.M.: Acoustical properties of speech as indicators of depression and suicidal risk. *Biomedical Engineering, IEEE Transactions on* **47**(7), 829–837 (2000)
25. Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L.V., Bardram, J.E.: Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 133–142. ACM (2013)
26. Hawthorne, G.: Measuring social isolation in older adults: development and initial validation of the friendship scale. *Social Indicators Research* **77**(3), 521–548 (2006)
27. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* **2**, 102–138 (1999)
28. Kasckow, J., Zickmund, S., Rotondi, A., Mrkva, A., Gurklis, J., Chinman, M., Fox, L., Loganathan, M., Hanusa, B., Haas, G.: Development of telehealth dialogues for monitoring suicidal patients with schizophrenia: Consumer feedback. *Community mental health journal* pp. 1–4 (2013)
29. Kirmayer, L.J., Robbins, J.M., Dworkind, M., Yaffe, M.J.: Somatization and the recognition of depression and anxiety in primary care. *The American journal of psychiatry* (1993)
30. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*, vol. 14, pp. 1137–1145 (1995)
31. Kotsiantis, S., Pintelas, P.: Predicting students marks in hellenic open university. In: *Advanced Learning Technologies, 2005. ICAALT 2005. Fifth IEEE International Conference on*, pp. 664–668 (2005). doi:10.1109/ICALT.2005.223
32. Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9. *Journal of general internal medicine* **16**(9), 606–613 (2001)
33. Lane, N.D., Mohammad, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., Campbell, A.: Bewell: A smartphone application to monitor, model and promote wellbeing. In: *Proc. of PervasiveHealth* (2011)
34. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: Stressense: Detecting stress in unconstrained acoustic environments using smartphones. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 351–360. ACM (2012)
35. Lu, H., Yang, J., Liu, Z., Lane, N.D., Choudhury, T., Campbell, A.T.: The jigsaw continuous sensing engine for mobile phone applications. In: *Proc. of SenSys* (2010)
36. Madan, A., Cebrian, M., Lazer, D., Pentland, A.: Social sensing for epidemiological behavior change. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 291–300. ACM (2010)
37. Martinez, D.: Predicting student outcomes using discriminant function analysis. (2001)

38. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: Proc. of SenSys (2008)
39. Pollak, J.P., Adams, P., Gay, G.: PAM: a photographic affect meter for frequent, in situ measurement of affect. In: Proc. of SIGCHI (2011)
40. Puiatti, A., Mudda, S., Giordano, S., Mayora, O.: Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. In: Proc. of EMBC (2011)
41. Rabbi, M., Ali, S., Choudhury, T., Berke, E.: Passive and in-situ assessment of mental and physical well-being using mobile sensors. In: Proc. of UbiComp (2011)
42. Rachuri, K.K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas, A.: Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 281–290 (2010)
43. Romero, C., Espejo, P.G., Zafra, A., Romero, J.R., Ventura, S.: Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education* **21**(1), 135–146 (2013)
44. Russell, D.W.: UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of personality assessment* **66**(1), 20–40 (1996)
45. Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* **4**, 1–32 (2008)
46. Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., Appleton, J.: Predicting student risks through longitudinal analysis. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 1544–1552. ACM, New York, NY, USA (2014). doi:10.1145/2623330.2623355. URL <http://doi.acm.org/10.1145/2623330.2623355>
47. Taylor, S.E., Welch, W.T., Kim, H.S., Sherman, D.K.: Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological Science* **18**(9), 831–837 (2007)
48. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), pp. 267–288 (1996). URL <http://www.jstor.org/stable/2346178>
49. Trockel, M.T., Barnes, M.D., Egget, D.L.: Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors. *Journal of American college health* **49**(3), 125–131 (2000)
50. Tudor-Locke, C., Sisson, S.B., Collova, T., Lee, S.M., Swan, P.D.: Pedometer-determined step count guidelines for classifying walking intensity in a young ostensibly healthy population. *Canadian Journal of Applied Physiology* **30**(6), 666–676 (2005)
51. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, pp. 3–14. ACM, New York, NY, USA (2014). doi:10.1145/2632048.2632054. URL <http://doi.acm.org/10.1145/2632048.2632054>
52. Wang, R., Harari, G., Hao, P., Zhou, X., Campbell, A.T.: Smartgpa: How smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pp. 295–306. ACM, New York, NY, USA (2015). doi:10.1145/2750858.2804251. URL <http://doi.acm.org/10.1145/2750858.2804251>
53. Watanabe, J.i., Matsuda, S., Yano, K.: Using wearable sensor badges to improve scholastic performance. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, pp. 139–142. ACM (2013)

54. Watanabe, J.I., Yano, K., Matsuda, S.: Relationship between physical behaviors of students and their scholastic performance. In: Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), pp. 170–177. IEEE (2013)
55. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology* **54**(6), 1063 (1988)
56. Xu, R., Wunsch, D., et al.: Survey of clustering algorithms. *Neural Networks, IEEE Transactions on* **16**(3), 645–678 (2005)
57. Zafra, A., Romero, C., Ventura, S.: Multiple instance learning for classifying students in learning management systems. *Expert Systems with Applications* **38**(12), 15,020–15,031 (2011). doi:<http://dx.doi.org/10.1016/j.eswa.2011.05.044>. URL <http://www.sciencedirect.com/science/article/pii/S0957417411008281>

Mobile Health

Sensors, Analytic Methods, and Applications

Rehg, J.M.; Murphy, S.A.; Kumar, S. (Eds.)

2017, XL, 542 p. 128 illus., 100 illus. in color.,

Hardcover

ISBN: 978-3-319-51393-5