Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning

RUNNING TITLE: Personal Sensing for Mental Health

David C. Mohr¹, Mi Zhang², Stephen M. Schueller¹

IN PRESS, ANNUAL REVIEW OF CLINICAL PSYCHOLOGY
DO NOT DISTRIBUTE

1. Center for Behavioral Intervention Technologies, Department of Preventive Medicine, Northwestern University
2. Department of Electrical and Computer Engineering, Michigan State University

Corresponding Author
David C. Mohr, Ph.D.
Center for Behavioral Intervention Technologies
750 N. Lakeshore Dr., 10th Floor
Chicago, IL 60611
Email: d-mohr@northwestern.edu

Stephen Schueller
Email: schueller@northwestern.edu

Mi Zhang
Email: mizhang@egr.msu.edu

Word count: 9,188
# TABLE OF CONTENTS

1. INTRODUCTION

2. FROM DATA TO KNOWLEDGE: A HIERARCHICAL MODEL
   2.1 Raw Sensor Data
   2.2 Feature Extraction: Data to Information
   2.3 Behavioral Markers: Information to Knowledge
   2.4 Clinical Targets

3. REVIEW OF PERSONAL SENSING RESEARCH
   3.1 Mobile Phones
   3.2 Other Devices and Platforms

4. METHODS
   4.1 Supervised Learning
   4.2 Unsupervised Learning
   4.3 Semi-Supervised Learning
   4.4 Deep Learning

5. CURRENT CHALLENGES IN PERSONAL SENSING
   5.1 Study Quality and Reproducibility
   5.2 The Curse of Variability
   5.3 The Unknown Expiration Date
   5.4 Balancing Accuracy and Invisibility
   5.5 The Certainty of Uncertainty
   5.6 Privacy, Ethics and the Naked Truth

6. POTENTIAL APPLICATION FOR PERSONAL SENSING
6.1 Integration into existing models of care

6.2 Behavioral Intervention Technologies

6.3 Epidemiology

7. SUMMARY AND CONCLUSIONS

Key words

mental health
mHealth
machine learning
pervasive health
wearables
sensors
ABSTRACT (150 words)

Sensors in everyday devices such as our phones, wearables, and computers leave a stream of digital traces. Personal sensing refers to the collection and analysis of data from sensors embedded in the context of daily life with the aim of identifying human behaviors, thoughts, feelings, and traits. This paper provides a critical review of personal sensing research related to mental health, focused principally smartphones, but also including studies of wearables, social media, and computers. We provide a layered, hierarchical model for translating raw sensor data into markers of behaviors and states related to mental health. Research methods as well as challenges, including privacy and problems of dimensionality are discussed. While personal sensing is still in its infancy, it holds great promise as a method for mental health research, and a clinical tool, monitoring at-risk populations and providing the foundation for the next generation of mHealth interventions.
1. INTRODUCTION

This paper is the story of the collision of several innovations – ubiquitous sensing, big data, and mHealth – and their potential to revolutionize mental health research and treatment. A sensor is any device that detects and measures a physical property. Sensors are as old as civilization itself. The Sumerians developed scales, which are essentially weight sensors, some 9000 years ago, and we have continued to develop new sensors ever since. The use of sensors to measure physical properties for the purpose of understanding psychological states, or psychophysiology, has long been a core discipline within psychology. Advances in sensor technology have accelerated throughout the last decades, with sensors becoming smaller, lighter, and more accurate. Furthermore, they have become increasingly ubiquitous and embedded into networks such that they can provide vast amounts of data almost anywhere and nearly instantaneously.

Today, people are measured continuously by sensors. Many sensors are embedded in mobile phones, measuring location, movement, communication/social interaction, light, sound, digital devices in the area, and more. Smartwatches and wearable devices containing onboard sensors that track activity and physiological functions are increasingly popular. People leave digital traces when they make credit card purchases, send a tweet, or visit a website. This digital exhaust produced by sensors has rich information about people’s behavior, and potentially their beliefs, emotions, and ultimately mental health.

Various terms have been used to describe the utilization of ubiquitous sensor data to estimate behaviors such as reality mining (Eagle & Pentland 2006), personal informatics (Li et al 2010), digital phenotyping (Jain et al 2015, Torous et al 2016), and personal sensing (Klasnja et al 2009). We use the term personal sensing as it is easily understood and conveys the intimacy of
the information. In this paper we provide an overarching model for personal sensing, review the literature on the use of sensors to detect mental health conditions and related behavioral markers, provide an overview of methods, and describe some of the grand challenges and opportunities in this emerging field.

2. FROM DATA TO KNOWLEDGE: A HIERARCHICAL MODEL

The goal of personal sensing applied to mental health is to convert the potentially large amount of raw sensor data into meaningful information related to behaviors, thoughts, emotions (for simplicity, in this paper we refer to these collectively as “behavioral markers”), and clinical states and disorders. While there are many approaches to sensemaking, we present here a layered, hierarchical sensemaking framework, as this illustrates a number of processes and issues. In this framework raw sensor data is captured and converted into “features” that contain information. These features can then be used to define behavioral markers often through machine learning. In the end, the entire set of features and behavioral markers can be used to identify clinical states such as diagnosis of a disorder. While some methods, such as deep learning (discussed further below), do not necessarily require these steps, we believe this framework is useful both because it is more likely viable in most academic research contexts, and because it illustrates several important issues in sensemaking. We will use a simplified version of a mobile phone sensing platform for common mental health problems as an example, depicted in Figure 1, however they could include data from any source.

2.1 Raw Sensor Data

The green boxes at the bottom represent the inputs into the sensing platform in the form of raw phone sensor data. For the most part, unprocessed raw sensor data does not contain the
sufficient information for the inferences we aim to make.

2.2 Feature Extraction: Data to Information

To add information, raw sensor data must be transformed into features. Features are constructs measured by and proximal to the sensor data. Features are depicted by the yellow boxes in the second layer of Figure 1. This is arguably the most important step in sensemaking (Bengio et al 2013). There are a number of ways to construct and extract features. One common approach is to use domain expertise or brainstorming to inject human intelligence for feature construction. For example, raw phone usage data may be of minimal value. If you are interested in in-phone communication, relevant features might be the number and duration of incoming calls or SMS, number and duration of outgoing calls and SMS, number of missed calls, and ratios of these features. In addition, features can also be extracted statistically using algorithms such as slow feature analysis and stacked autoencoders (Vincent et al 2010, Wiskott & Sejnowski 2002) that can automatically discover new feature representations. Finally, some features estimate “observable” states using machine learning. For example, bed/waketime can be estimated using a number of sensors and features related to light, sound, and phone use (Zhenyu et al 2013).

2.3 Behavioral Markers: Information to Knowledge

Behavioral markers are higher-level features reflecting behaviors, cognitions, and emotions that are measured using low-level features and sensor data. This is similar to notion of latent constructs in psychological methodology. Some examples of potential behavioral markers are represented in Figure 1 by the blue boxes. Behavioral markers are most commonly developed using machine learning and data mining methods to uncover which features and sensor data are useful in detecting the marker. For example, a behavioral marker for circadian sleep rhythm
might include features such as bed/waketime, sleep duration, and phone usage. Sleep quality might include ambient sound features, but may also include bed/wake time (Abdullah et al 2014). Furthermore, the accuracy of such features may be enriched by including additional “helper” features such age (older people use phones differently from younger people) or whether it is a workday or non-workday.

2.4 Clinical Targets

One would not attempt to diagnose a mental health disorder on the basis of one or two questions about symptoms (although one might use them for screening purposes). Similarly, as will be discussed below, limited sets of features have been only modestly successful at predicting clinical targets. We expect that clinical targets will be better predicted by applying machine learning methods to a larger number of behavioral markers and features. We note that these may not have a one-to-one correspondence to symptoms used to diagnose disorders. Some symptoms may simply not be detectable and personal sensing may uncover other predictors that have not been considered to date.

3. REVIEW OF PERSONAL SENSING RESEARCH

Most work on personal sensing for mental health has used mobile phone sensors. We will therefore first review this work before reviewing work from other areas, including wearables, social media, computers.

3.1 Mobile Phones

Mobile phones are commonly used for research because they are widely owned, 64% of Americans own a smartphone, up from 35% in 2011 (Smith 2015). Furthermore, people keep their phones on or near them and use them frequently. People check their phones on average 46 times per day, and for younger people that figure is 85 times per day (Andrews et al 2015,
Eadicicco 2015). The phone also has an increasingly large number of embedded sensors. Here we focus on behavioral markers in three areas related to mental health: sleep and social context, which are potentially more observable, and mood and stress, which are internal states.

3.1.1 Behavioral Markers

Sleep

Sleep disturbance is a common symptom occurring across many mental health conditions (Sivertsen et al 2009, Taylor et al 2005). Those disturbances can be reflected by patients' sleep periods (i.e., when and how long a person sleeps) and sleep quality (i.e., how well a person sleeps). By leveraging the built-in sensors, a number of smartphone-based sensing systems have been developed to passively monitor sleep periods. Several groups have shown that sleep duration can be estimated with around 90% accuracy, without asking the user to do anything special with the phone, using data from a number of sensors, such as accelerometer, microphone, ambient light sensor, screen proximity sensor, running process, battery state, and display screen state (Chen et al 2013, Min et al 2014). Among heavy phone users, such as undergraduate students, sleep periods can be detected simply by observing phone screen lock and unlock events, which is less of a drain on the phone battery than other methods (Abdullah et al 2014). Sleep period markers can then be used to create circadian-aware systems. For example, non-workday sleep duration can be used to estimate a person’s chronotype (e.g. morning lark vs. night owl), and changes in sleep patterns across work and non-workdays can identify social jet lag, which is the difference between a person’s biological sleep rhythm and external requirements (Abdullah et al 2014, Murnane et al 2015). Such sleep period markers have also been correlated with depressive symptom severity (Wang et al 2014).

Sleep quality can also be effectively inferred by smartphone sensors. For example,
common events that interfere with sleep quality, such as body movement, coughing and snoring, and ambient noise can be reliably detected using a smartphone’s microphone, when the phone is kept in the user’s room. Acoustic features have been associated with both short-term (i.e., one-night) and long-term sleep quality measured using actigraphy and self-report (Hao et al 2013). A number of studies have used multimodal sensing schemes, including the accelerometer, microphone, light sensor, screen proximity sensor, running process, battery state, and display screen state, to infer sleep stages and sleep quality (Gu et al 2014, Min et al 2014).

Social Context

We focus here on phone-based research, and discuss social media network work below. Large population-based phone data sets can provide dynamic information on individual movement and proximity to others that can be used to calculate people’s proximity in a social network. Patterns of movement and of co-location can be used to infer relationships and predict new social ties in the future (Hsieh & Li 2014, Pham et al 2013, Wang et al 2011).

On a smaller scale, in a classic study, Eagle was able to identify friends and non-friends with a high degree of accuracy using Bluetooth sensors (Eagle & Pentland 2006, Eagle et al 2009), which can detect other Bluetooth enabled devices up to 15 meters away. People in proximity to one another only during work hours were more likely to be colleagues than friends, while proximity during the evening or weekends was an indicator of friendship. Relational status can then be used to identify other psychological targets. For example, calling friends during work hours was associated with lower job satisfaction (Eagle et al 2009). While such methods hold promise, their application is limited by the small percentage of people who leave their Bluetooth sensors in discoverable mode. Nevertheless, Eagle’s work underscores the utility of co-location and time as important features indicating the nature of relationships.
Other forms of social sensing have used remote communication tools within the phone, including calls and SMS. Contact lists (address books) within a person’s phone can contain information about relationships. For example, contact fields sometimes include family role (e.g. Aunt Julie), relationship context (e.g., Kaitlyn (Peg’s friend)), phone type (e.g., Mom at home), or an honorific (Mrs., Mr., Dr., etc.), which can be mined to infer relationships (Wiese et al 2014). However, contact lists are vulnerable to idiosyncratic labeling methods, and tell us little about the frequency of contact.

Patterns of time, frequency, and regularity of incoming and outgoing calls and SMS have also been able to classify a person’s contacts into a relationship domain (family, friend, or work) with more than 90% accuracy (Min et al 2013). For example, longer calls were associated with family, work was characterized by fewer weekend calls and lower likelihood of SMS, while friend and social contacts were characterized by more SMS during the week. The strength of social ties can also be estimated to some degree, with higher levels of in-phone communication frequency, call duration, and communication initiated by the phone owner being associated with stronger relationship (Wiese et al 2015). However, this signal is noisy, because low levels of communication do not necessarily mean weak ties (we speak infrequently with some people who are very close to us) as much communication may occur outside of the phone such as face-to-face or other media, and increasing other applications such as Snapchat and WhatsApp.

*Mood and stress*

Mood and stress are internal states that are likely more distal from the sensors and features normally used in personal sensing. A number of studies have attempted to leverage a broad array of built-in mobile phone sensors to predict mood (LiKamWa et al 2013, Ma et al
2012, Madan et al 2010). In the earliest study, Madan found decreased calls, SMS, Bluetooth detected contacts, and location entropy (a measure of the temporal dispersion of locations) were strongly related to feeling sad and stressed among students, as measured by daily EMAs (Madan et al 2010). MoodScope (LiKamWa et al 2013) inferred mood, labeled by daily EMA from 32 participants over two months. The number and length of communications (calls, texts, and emails), number and app usage patterns, web browser history, and location could estimate a user’s daily mood average with an initial accuracy of 66%, which gradually improved to 93% after a two-month personalized training period. Similarly, using location, motion detectors, light, and ambient sound, Ma achieved approximately 50% accuracy for daily moods over a 30 day period with 15 participants (Ma et al 2012). A recent attempt to replicate the MoodScope findings on a cohort of 27 students failed to perform better than chance (Asselbergs et al 2016). The varying results and failure to replicate suggest that while a number of small studies have demonstrated the technical feasibility of sensing mood, these findings do not appear to generalize.

At least one study has attempted to detect stress using the “swipe,” “scroll,” and “text input” interactions with the phone (Ciman et al 2015). This work, based in part on literature showing that stress can be detected through computer mouse and keyboard interactions (see below), found that under laboratory conditions, features derived from a person’s “scroll,” “swipe,” “touch,” and “text input” interactions with the phone could differentiate a laboratory induced stressful state from a normal state. Whether real-world instances of these interactions provide a strong enough signal for stress remains an open question.

A large literature has demonstrated that affect and mood can be detected through the paralinguistic features of speech (Calvo & D'Mello 2010). StressSense (Lu et al 2012) is a
smartphone sensing system that uses the built-in microphone to capture human speech during social interactions to infer whether a user’s level of perceived stress by analyzing the paralinguistic information such as pitch and speaking rate. Under quasi-experimental conditions using a mock job interview and a marketing task, StressSense achieved 76-81% accuracy on stress identification. These findings were then extended to real-world evaluation. Following 7 participants over 10 days, the sensed stress marker correlated at $r=.59$ with self-reported stress (Adams et al. 2014). Another example, EmotionSense (Rachuri et al. 2010) proposed that audio-based emotion recognition could identify up to 14 different emotions clustered into five broader emotion groups (happy, sad, fear, anger, and neutral). In an initial proof-of-principle 10-day study involving 18 participants, the distribution of the emotions detected through EmotionSense generally reflected the self-reports by the participants.

The detection of mood or subjective stress is likely a challenge for many commonly available sensors in the phone. Given the history of paralinguistic voice features predicting mood, the microphone would seem promising from an analytic perspective but may pose challenges from a logistical and ethical perspective in acquiring samples of sufficient quantity and quality. This illustrates a disconnect that sometimes occurs between technical and laboratory proof-of-concept and real-world feasibility.

### 3.1.2 Clinical Disorders

**Depression**

Early work on using smartphones for personal sensing began examining depression. Madan followed 70 undergraduates living in a residence hall and found that decreases in total communication were associated with greater depression (Madan et al. 2010). Depression, however, was assessed only using a single item. A second study, StudentLife, used the PHQ-9 to
assess depression, among 48 students over 10 weeks (Wang et al 2014). Similar to the Madan study, conversation frequency and duration using the microphone, as well as co-location with other students detected using GPS and Bluetooth were significantly related to depression. In addition, depression was associated with a previously developed sleep duration classifier (Chen et al 2013). The relationship between sensed social contact and depression was also observed in a non-student population using elderly people living in a retirement community (Berke et al 2011).

These relationships are perhaps unsurprising, given sleep disruption is a symptom of depression, and social withdrawal and avoidance are significant factors related to common mental health problems such as depression and anxiety (Hames et al 2013). The fact that sleep, social withdrawal, and anxiety were inferred through smartphone sensors demonstrates the utility of the hierarchical model displayed in Figure 1.

While sensing has the potential to automate detection of behaviors we know are related to disease states, it also has the potential to uncover new information and understanding. A case in point is the emerging work on the relationship of GPS data to depression. The first study in this area, using two weeks of data from 28 participants, found a number of GPS-derived location features associated with depression (Saeb et al 2015). The number of places a person visited was not related to depression, however, location entropy (the variability in time spent in different locations), was such that the more time clustered around a few locations, the more likely the person was to be depressed, while more equal time distributions was related to lower depression scores. A feature measuring periodicity, or circadian rhythm of movement through geographic space was particularly strongly related to depression, suggesting that disruption in the regularity in movement was associated with greater severity of depressive symptoms. These findings were then replicated in the StudentLife data set described above (Saeb et al Under Review). A third
study, using somewhat different methods, similarly found that similar GPS features could estimate depression (Canzian & Musolesi 2015).

This general relationship between mobility and depression has been explored in more detail. For example, the relationship between GPS features and depression is stronger on non-workdays than it is on workdays when much movement is driven by social expectations (Saeb et al Under Review). This suggests that distinguishing between times when behaviors are more under the individual’s control versus when they are not may identify features that can be useful in increasing the accuracy of models. GPS features appear to predict depression many weeks in advance, while the relationship between depression and subsequent GPS features degrades quickly over time, suggesting that human mobility may be an early warning signal for depression.

Bipolar Disorder

The MONARCA project was a pioneer in smartphone-based behavior monitoring technologies for mental health (Grünerbl et al 2015). MONARCA leveraged a variety of phone sensors to detect the mental states of patients as well as changes in mental states. To validate the effectiveness of their smartphone-based sensing system, the MONARCA team conducted a series of real-world studies to develop were conducted among bipolar patients from a rural psychiatric hospital in Austria. Based on 12 patients followed over 12 weeks, accelerometry, location, or fused accelerometry/location features produced clinical state (depression/manic) recognition accuracy of 72-81% and state change detection with a precision/recall of 96%/94% (Grünerbl et al 2014). By fusing phone call features and paralinguistic information, a state recognition accuracy of 76% as well as a precision/recall of 97%/97% for state change detection were achieved (Grünerbl et al 2015). Another analysis of 18 patients over 5 months indicated
that app use can identify stress and mood (Alvarez-Lozano et al 2014). For example, higher use of social and entertainment apps was associated with lower stress and irritability.

Work has also examined the potential for GPS features, originally developed for depression (Saeb et al 2015), to detect depressive episodes among bipolar patients. The same features, including entropy, circadian rhythm, remain strongly related to severity of depression in this population (Palmius et al Under Review). Furthermore, when combined, these features could classify depressed from non-depressed states with 85% accuracy. This underscores the potential utility of features across diagnoses when examining similar states.

Schizophrenia

Work on sensing in schizophrenia has begun just recently. A survey of patients with schizophrenia suggested that most are comfortable using a phone with sensing, and are interested in potentially receiving feedback and suggestions from such a system, although a minority voiced concerns that it might upset them or were concerned about loss of privacy (Ben-Zeev et al 2016). In a first study, 34 patients with schizophrenia were provided a phone over 2-8.5 months that collected a variety of sensor data. Personalized models used a number of features to predict EMA responses. For example, changes in physical activity, detected conversations, and later bedtimes were associated with self-reported worry that someone is intending to harm the participant was associated, and self-reported auditory and visual hallucinations (Ben-Zeev Under Review).

3.2 Other Devices and Platforms

We have described life sensing using mobile phones as this is the most ubiquitous personal sensing platform, harnessing data from people’s lives with little to no ongoing effort or actions on the part of the user. However, many other sources of data exist, which have their own
strengths and weaknesses. We review below data from wearable, social media, and computers.

3.2.1 Wearables

Wearable devices (AKA wearables) are sensor-enabled technologies designed to be worn for specific purposes, most commonly health and fitness. These devices, such as Fitbit and Jawbone, track activities continuously, for example, how many steps people take, how many miles they run, and how long they sleep. Wearables, which use sensors designed for their specific targets, and which are intended to worn in specified and consistent manner (e.g. on the wrist or clipped to the belt), may provide data that is of significantly higher quality than that provided by phones, which are not designed specifically for health tracking. However, the increase in data quality may be offset by other drawbacks. Wearables are less prevalent than smartphones with only 19% ownership among Americans (Ricker 2015). Their use is higher among those already motivated to keep a watchful eye over their health and many people abandon using them soon after purchase (Piwek et al 2016).

The most widely used sensor in wearables is the accelerometer. Accelerometer-based wearable devices have been developed for physical exercise tracking (Choudhury et al 2008), fall detection (Li et al 2009), and activities of daily living (ADL) monitoring (Spenkelink et al 2002). In a large study of 2,862 participants, greater levels of accelerometry-based of physical activity were strongly associated with decreased rates of depression (Vallance et al 2011).

Increasingly wearable devices are including a broader range of sensors that can measure variables that are useful for mental health researchers, such as skin conductance and heart rate. For example, investigators have noted that greater asymmetries in skin conductance amplitude on the left and right side of the body are an indicator of emotional arousal (Picard et al 2016). Many of these sensors are now available in smartwatches, which attempt to leverage a behavioral
and cultural pattern to avoid the problem of abandonment seen with other wearables. As of this writing, it remains to be seen if smartwatches will attain the ubiquity enjoyed by smartphones.

Wearables are being developed that are dedicated to behaviors that have been difficult to sense. Eating and appetite, for example, are often disrupted in mental health conditions, but are difficult to detect through commonly available sensors. Two specific methods, gestures and sound, may be particular useful for these behaviors. Because eating and drinking activities normally involve repetitive wrist movements and rotations, wrist-worn wearables that include an accelerometer and gyroscope have shown promise in capturing eating and drinking activities (Edison et al 2015, Sen et al 2015). Eating and drinking may also produce idiosyncratic sounds through chewing and swallowing. A microphone attached at the neck can classify sounds produced by eating and drinking with reasonable accuracy (Kalantarian et al 2015, Rahman et al 2014, Yatani & Truong 2012).

3.2.2 Social media

With 65% of Americans using social media in 2015 (Perrin 2015), platforms, such as Facebook and Twitter, have become common platforms where people share their opinions, feelings, and daily experiences. The field of psycholinguistics has long demonstrated that linguistic analysis of speech can be used for diagnostic classifications (Oxman et al 1982, Rude et al 2004). Thus, social media postings, which consist largely of language, are a potential source of information about mental health, as well as people’s thoughts and feelings, related to those conditions. Using a large dataset of more than 28,000 Facebook users who completed a personality survey, Schwartz et al (2014) found that features generated from posts were modestly related to depression severity. Themes related to depression include depressed mood, hopelessness, hopelessness and helplessness, symptoms, relationships and loneliness, hostility,
and suicidality. Similarly De Choudhary (2013b) found that depressed Twitter users can be distinguished from non-depressed based on later posting time, less frequent posting, greater use of first person pronouns, and greater disclosure about symptoms, treatment and relationships. Furthermore, the development of a future depressive episode could be predicted with 70% accuracy. In a large sample of Twitter users, rates of depression were consistent with geographic, demographic, and seasonal patterns reported by the Centers for Disease Control (De Choudhury et al 2013a).

Social media likely will be helpful at identifying behavioral markers that are strongly related to cognitive and motivational factors, which are difficult to evaluate through non-verbal sensors. For example, language features from Facebook posts have shown modest but consistent correlations with Big 5 personality factors (Park et al 2015). Twitter derived features related to suicidal ideation have been shown to correlate strongly with rates of completed suicides from the CDC (Jashinsky et al 2014). Suicidal ideation has also been associated language used in social media showing heightened self-attentional focus, poor linguistic coherence and coordination with the community, reduced social engagement, and manifestations of hopelessness, anxiety, impulsiveness and loneliness (De Choudhury et al 2016). Thus, language generated naturalistically through social media may be a useful tool in sensing mental health conditions, and may be particularly well suited for behavioral markers that involve cognitive or motivational states that are beyond the reach of nonverbal sensors.

3.2.3 Computers

Many people spend a considerable amount of their lives at computers and many interactions still take place through the mouse and keyboard. A number of studies have examined whether mouse movements and keyboard taps can provide information on a person’s
mental state. An early study explored a broad range of possible features derived from mouse movements to predict experimentally induced emotions (Maehr 2008). While most features showed no relationships to emotions, “motion breaks”, or discontinuities in movement, were related to overall arousal and discrete emotions such as disgust and anger. Motion breaks resemble pause features observed in speech that have been related to stress and it is possible that this is a general behavioral pattern, observable across multiple channels, when one is stressed. Another study explored whether muscle tension would change the dynamics of the movement (resonant frequency and damping ratio), and thus be an observable correlate of stress. Data collected in a laboratory setting demonstrated that simple models of arm-hand dynamics applied to mouse motions were strongly related to concurrently collected physiological measures of stress and arousal (Sun et al 2014). This signal remained strong across a variety of mouse tasks including clicking, dragging, and steering.

Similar affective inferences have been made from keyboard activity. One study tracking everyday computer use for one month among 12 participants found promising accuracy (70-88%) for self-reported emotion (Epp et al 2011). Models were trained using a decision-tree classifier and features derived from short key sequences, such as duration of and latency between keystrokes, to predict common discrete emotional states. While these results were encouraging, classification rates represented only a modest gain over baseline classification.

3.2.4 Additional Sources of Data

Phones, wearables, social media, and computers are far from the only technologies that produce digital traces. Other potential streams of data include purchasing behavior, browsing history, or productivity apps such as calendars and email. Furthermore, social context could be better understood by making use of Google maps or other repositories of images of
environments. Such images can be mined to determine environmental factors that affect mental health and well-being (e.g. amount of green space or number of trees in a neighborhood, or cleanliness or number of tagged surfaces). We have not discussed these, primarily because they have not yet been investigated in relationship to mental health. Another rapidly expanding area is “ambient intelligence” where sensors are installed on everyday objects and in living environments to sense people’s movements, gestures, habits, and intentions and respond to needs in a seamless and non-intrusive manner (Acampora et al 2013). Indeed, such ambient systems have the potential to provide visibility into the most intimate spheres of a person’s life.

4. METHODS

The field of personal sensing is very young, with almost all of the research occurring in the last few years. Most of the studies have been conducted by computer science and engineering groups which have research models that are very different from those commonly used in the clinical and behavioral sciences (Intille 2013). As we in the behavioral sciences begin work in this area, it is important to understand the fundamental differences between these methods and those with which we are more familiar.

First, engineering and computer science research is typically exploratory in nature, focused on solving a problem. In the area of personal sensing, computer scientists tend to collect as much data as possible, using data mining methods to develop classification algorithms. While these analytic methods employ techniques such as cross validation to avoid overfitting in the models, these methods are nonetheless quite different from commonly used clinical methods, which come from a positivist tradition, and are hypothesis driven and confirmatory rather than exploratory. Said a different way, clinical scientists tend to design a study to “test” an answer to a question. Engineers tend to design a study to “find” an answer to a question.
Engineers and computer scientists, in their quest for novel solutions, tend to have a higher tolerance for risk and change in research than do clinical scientists. Clinical scientists place a much higher value on eliminating as many threats to internal validity as possible, and have a low tolerance for methods that might limit confidence in the results. As such clinical scientists aim to avoid Type I errors or spurious positive findings while engineers and computer scientists see Type II errors, which would lead to overlooking a potentially novel and useful solution, as the greater threat.

Finally, engineering and computer science are frequently focused on proof-of-principle, while clinical scientists value generalizability. It is perhaps only a slight exaggeration to say that computer scientists in personal sensing are asking “Does this work at all?” while clinical scientists want to ask, “Will this work for a population under all circumstances?” Thus, most of the studies in personal sensing have been small, generally with sample sizes of 7-30 participants, commonly using convenience samples consisting of college students (the social media studies are the exception, using very large data sets). It is not uncommon for comparatively large percentages of enrolled participants, sometimes on the order of half the sample, to be excluded from analyses due to any number of problems in data acquisition or data quality. Thus, a clinical scientist might see these as offering little assurance that the findings might extend outside of that research context. On the other hand, engineers and computer scientists have demonstrated evidence that a novel solution may have value.

The main analytic method used for personal sensing, is machine learning (Bishop 2006). The goal of machine learning is to identify potentially complex relationships among data, and to use the identified relationships to make predictions on new data. We review here three commonly used machine learning analytic methods: supervised learning, unsupervised learning,
and semi-supervised learning as well as a new trend in machine learning called deep learning in the context of personal sensing.

4.1 Supervised Learning

Supervised learning is a category of algorithms in machine learning that aims to learn a function that maps data to “labels” provided by a set of training samples. A label in machine learning refers to that which is being predicted, similar to a dependent variable in statistics. In personal sensing, labels are often user self-reports. A “training sample” is a pair consisting of a data instance and its label. Like a teacher supervising learning in a classroom, the labels “supervise” the learning process, which occurs through training samples. The learned mapping function is then applied to data in the absence of labels to predict their labels. If labels are categorical values, the supervised learning algorithms are called “classification algorithms,” and the mapping function is referred to as a classifier. Learning algorithms for continuous values are called “regression algorithms,” and the mapping function is called a regression function.

Classification is the most commonly used supervised learning method. For example, activity recognition can be formulated as a classification problem where sensor data are mapped to different activity labels such as walking, running, sleeping, etc. There are two families of classification algorithms: generative algorithms and discriminative algorithms. Generative algorithms learn a classifier of the joint probability of the data instances and their labels, and then calculate the posterior probability by applying Bayes theorem to predict labels of new data instances (Ng & Jordan 2002). Naive Bayes, hidden Markov models, and Gaussian mixture models are some of the most commonly used generative algorithms. In contrast, discriminative algorithms build a model to describe the boundaries that separate different labels. Examples of discriminative algorithms include logistic regression, support vector machine, and conditional
random fields. Generative and discriminative algorithms have unique strengths and weaknesses. In practice, the classification performance of discriminative algorithms tends to be better than generative algorithms (Bishop & Lasserre 2007). However, generative algorithms can identify data that come from a new label that is not included in the training samples, for example identifying new activities of a user (e.g., yoga) that are not included in the training samples.

4.2 Unsupervised Learning

The goal of unsupervised learning is to find hidden structure within the data. In unsupervised learning, the training samples do not have labels and only contain data instances. There are three families of unsupervised learning algorithms: clustering, anomaly detection, and dimensionality reduction, each aiming to identify different structure within the data. Clustering algorithms (such as K-means and hierarchical clustering) aim to divide data instances into separate clusters such that data in the same cluster are similar while data in different clusters are dissimilar. Anomaly detection algorithms (such as one-class support vector machines) aims to identify the few instances that are very different from the majority of the data. Finally, dimensionality reduction algorithms (such as feature selection and Principal Component Analysis) aim to remove the multi-collinearity and retain the most important information of the data to avoid the effects of the “Curse of Dimensionality” (Bishop 2006), thereby improving the generalization performance of machine learning models.

In personal sensing, unsupervised learning algorithms are often used to preprocess sensor data before using supervised methods for further processing. For example, clustering algorithms have been applied to GPS data (i.e., pairs of latitudes and longitudes) to create heat map and to find points of interests of the user (e.g., home, work place) (Saeb et al 2015). Anomaly detection algorithms have been used to detect changes of mental states of bipolar patients so that just-in-
time intervention can be delivered (Grünerbl et al 2014). Finally, dimensionality reduction algorithms have been applied to identify the most important behavioral markers to best predict the mental states of depressive patients (Saeb et al 2015).

4.3 Semi-Supervised Learning

As its name implies, semi-supervised learning is in the middle of supervised and unsupervised learning. It uses training samples that contain both labeled and unlabeled data to achieve better performance than could be achieved by simply using supervised learning trained on the labeled data (Zhu & Goldberg 2009). Semi-supervised learning is very practical for personal sensing, where there is a large ratio of unlabeled to labeled data. For example, it would be either burdensome, expensive, and time-consuming, if not impossible, to label such every minute of GPS or accelerometry data collected throughout a day (Chapelle et al 2006). Semi-supervised learning addresses this problem by leveraging the intrinsic structure of the unlabeled data together with information provided by the labeled data.

Active learning is a special case of semi-supervised learning. Active learning algorithms query a user to provide additional labels when the algorithm detects a user’s behavior or state deviates from what has been trained before and thus, the algorithm is uncertain how to classify. Therefore, in contrast to supervised learning models which are static and cannot be updated after the training period, active learning is able to update users’ models after getting additional labels. In this way, it can provide evolving models that are adaptive to users’ changing behaviors and states.

Active learning is especially useful in generating personalized models from group models. Group models are algorithms that, once developed, are intended to run passively, with no input from the user (much like activity tracking devices). Personalized models, require user
labeling to create a model that is specific to the individual. Personalized models tend to perform better than group models but incur labeling burden. However, this labeling burden may be somewhat mitigated with hybrid models that are initiated with group models and optimized through user labeling via active learning. In particular, active learning can help derive personalized models with less labeling, as the algorithm only requests labels when they are needed (Settles 2010).

4.4 Deep Learning

In the past decade, a new trend in machine learning called deep learning has emerged (Schmidhuber 2015). Methods developed based on deep learning have dramatically improved the state-of-the-art and have beaten other machine learning methods in a wide range of applications such as identifying objects in images (Krizhevsky et al 2012), speech recognition (Hinton et al 2012), language translation (Sutskever et al 2014), understanding the genetic determinants of diseases (Xiong et al 2015), and predicting health status using electronic health records (Miotto et al 2016).

The success of deep learning is rooted in a revolutionary way of extracting features from data. It is well understood that the performance of machine learning methods largely depends on the choices of features (Bengio et al 2013), which traditionally has required considerable human effort and domain knowledge. Although these hand-engineered features exhibit great performance in small data sets, they do not generalize well to challenging problems involving large-scale data sets (LeCun et al 2015). In contrast, deep learning adopts a data-driven approach in which a general-purpose procedure automatically learns features from data, with no prior domain knowledge needed. These self-learned features are organized in a multi-level hierarchy, where higher-level features are defined from lower-level ones, similar to the layered, hierarchical
sense-making framework illustrated in Figure 1.

Deep learning may be vulnerable to overfitting at smaller sample sizes, often making traditional machine learning methods a better fit. However, once an adequate sample size is obtained, deep learning exhibits superior capability at capturing the intricate characteristics of data that traditional machine learning methods fail to capture. Therefore, deep learning achieves much better performance than other methods as the sample size increases. Furthermore, although self-learned multi-level features generated purely by machines may not be well understood by humans, it is possible that these features may uncover new understanding about the constructs we are measuring but may do so by increasing the complexity beyond human understanding.

5. CURRENT CHALLENGES IN PERSONAL SENSING

5.1 Study Quality and Reproducibility

A growing number of studies would appear to show replication of findings of studies (albeit using small, narrow samples) using machine learning methods that estimate behavioral markers such as mood, stress (LiKamWa et al 2013, Ma et al 2012, Madan et al 2010), and sleep (Abdullah et al 2014, Chen et al 2013, Min et al 2014) using a combination of phone sensor data and features. However, as computer science and engineering tend to value technical novelty over generalizability, studies that appear to address the same behavioral marker use different sensors, different sets of features, different methods of measuring the behavioral markers, and varying research designs, (e.g. giving people phones vs. using their own, varying periods of time, varying numbers of participants excluded, etc.). The machine learning methods used vary, and the results or weightings, particularly for group models, are not necessarily comparable across studies. In addition, it is unclear how many attempts have not published due to failure. The one replication study we are aware of, using nearly identical methods, was unable to reproduce the
very strong findings in the original paper on prediction of mood (Asselbergs et al 2016). Thus, studies examining the use of machine learning methods to estimate behavioral markers indicate that it is feasible under narrow conditions, however the reliability needed for clinical use has not been demonstrated.

Furthermore, the availability of easy-to-use tools for machine learning is expanding faster than the expertise, resulting in a growing number of publications with questionable methods. A recent review of papers using sensor data to detect disease states found that half used inappropriate cross-validation techniques, which greatly overestimate prediction accuracy (Saeb et al 2016). Furthermore, papers that used these inappropriate techniques were cited just as often as papers using proper techniques, suggesting poor quality information is having the same impact as high quality information. We note that while the papers cited in this review did not evidence these types of methodological problems, it would behoove the interested scientists to explore the field with a healthy mix of excitement and skepticism.

5.2 The Curse of Variability

As we move from narrow proof-of-concept studies to testing in broader populations, the sources of variability expand enormously, emanating from a variety of sources including data types, characteristics of people, and different environments. Sensors in phones vary from manufacturer to manufacturer, model to model, and version to version, affecting the data. People’s characteristics might impact the relationship between constructs or how they use the measurement devices. For example, age may be related to the number of social contacts, with older people having and wanting less contact, but may also be related to how social activity is measured with a phone (e.g. older people are more likely to call and less likely to text than younger people). Where people carry the phone (pocket, handbag, backpack) can can profoundly
affect the sensor data. Environment and seasonality represent additional important dimensions. For example, GPS and accelerometer data in winter will look different in Minneapolis relative to Miami.

When dimensionality is high, individual small studies are unlikely to be adequately powered to create reliable and generalizable classifiers for use in larger populations. Efforts such as the Personalized Medicine Initiative (Collins & Varmus 2015), which plans to enroll more than 1,000,000 people, may provide such opportunities, but it is unclear as we write this what data will be collected and how. An approach used in other fields with similar problems, such as genomics, is to pool data across studies. A challenge in pooling data is to find a scientifically valid balance between identifying uniform variables, which makes data pooling straightforward (e.g. exact same questions) but can be hard to implement, and using statistical methods to manage heterogeneity by providing similar, if not identical data points (Fortier et al 2011). The field of personal sensing in mental health is still young and small enough that some agreement on a core set of clinical assessment methods (EMA or self-report) may be possible, thereby providing uniform anchors to which the broad range of sensor data, evolving and changing over time and across research projects, could be tied.

5.3 The Unknown Expiration Date

Personal sensing algorithms will likely have shelf lives, which may be relatively short. As devices and sensors are updated, the associated raw data will change over time. Additionally, the way we use these devices and platforms changes as well. Just in the past few years, phone use has changed dramatically. We spend more time reading and watching movies on our phones, and communications have shifted away from calls and SMS to messaging apps and social media. Social media is becoming increasingly more visual relative to text based, and interfaces and
notification methods are changing when, how, and what people write. As people change how they use the devices that provide the data, machine learning algorithms will become increasingly inaccurate.

Google Flu Trends offers a high profile cautionary tale. Launched in 2008, mining flu-related search terms produced results that closely matched the CDC’s surveillance data and provided the information more rapidly (Butler 2013). The system was rolled out to 29 countries and extended to other diseases. It performed remarkably well… until it stopped working. How people conducted searches changed over time, rendering the algorithms ineffective. The changes in people’s search strategies were driven at least in part by Google’s own efforts to optimize search algorithms, which also altered the search recommendations provided to users, changing people’s search behaviors, and ultimately undermining Google Flu Trend models.

5.4 Balancing Accuracy and Invisibility

A common goal in personal sensing is to make acquisition of data and prediction of behavioral markers as unobtrusive, or invisible to the user, as possible. On the one hand, requiring user actions will likely result in abandonment of the tool by some percentage of the users. On the other hand, personalized, active learning models, which require user labeling, perform better than static group models. Active learning models also allow recalibration over time, potentially eliminating the shelf life problem.

Thus, users who provide a little bit of data would enjoy substantially higher predictive accuracy of models. Rather than thinking of a sensing platform as a technology that autonomously creates information, it may be more useful to think of the sensing platform as a social machine in which the quality of prediction is a shared endeavor. The ability to accurately predict a marker or phenotype depends upon the use of data, passively and actively collected,
from many other individuals who have come before. Providing labeled data back into the system can improve the accuracy that individual user, as well as for all subsequent users, harnessing the wisdom of the crowd, while contributing to the crowd.

5.5 The Certainty of Uncertainty

The output of any personal sensing system, even under the best of circumstances, will always have some degree of error and uncertainty. This error is always user-facing, affecting the quality of the experience. This raises several questions that can and should be considered from the early stages of research. First, how much uncertainty is acceptable, and how much accuracy is good enough (Lim & Dey 2011)? For example, if a system were designed to detect likely depression among general internal medicine patients, how many false positives would be acceptable to clinic staff? What levels of false negatives would be acceptable to a care system or to patients, and how could the effect of false negatives be mitigated? In addition, error can be shifted between false positives and false negatives, depending on where it can best be managed and produce the least harm. Or perhaps a metric can be used in a way the minimizes the effect of inaccuracy. For example, step counts from activity trackers may be inaccurate, however, to the degree they are consistent within a user, they can be used for day-to-day comparisons. Early stage research can explore the understanding and acceptability of error and uncertainty and how best to mitigate it (Kay et al 2015).

5.6 Privacy, Ethics and the Naked Truth

The use of passively collected digital data raises many issues of privacy and security, about which there is disagreement within the community of researchers and a lack of guidelines (Shilton & Sayles 2016). We review here some general themes and topics that are most relevant to leveraging passive data for mental health purposes.
The principle of privacy refers to ensuring that people have choice and control over the use of their own data, and some would argue, that they understand those choices (Shilton 2009). Security refers to the protections put in place to ensure that people’s choices are followed. People’s agreement to share their data usually revolves around two key concepts: trust and value. Trust refers to the notion that the use of data will be appropriate given a person’s wishes and expectations. Value refers to the benefit that is accrued to the user or society based on the use of data.

An important aspect of trust has traditionally been severance of the identify of the individual from the data provided from that individual, or de-identification. This is challenging as even a few pieces of information, such as gender, zip code, and birthdate, can identify most of the US population (Sweeney 2000). The data collected from devices may pose even greater risks to identification. GPS traces are the most personally identifying type of data; with only 4 spatio-temporal points, 95% of individuals can be identified (de Montjoye et al 2013). Various methods can help obfuscate location data, however, none of these successfully preserve privacy while retaining its usefulness (Brush et al 2010).

Privacy management needs to give participants as much control over their data as possible (Shilton 2009). Participants should be informed what the data might reveal about them, how long the data will be used, who will be using it, and why. Data management tools can be designed that provide tools to help people manage their data, including the ability to define acceptable use, limit data access, delete data, or revoke consent altogether.

Greater openness, more transparency, and better methods to share data are desirable for several reasons including improving the quality of the scientific literature, providing opportunities to replicate findings, and creating tools that are valid, reliable, and generalizable.
As standards and best practices evolve around participant privacy, the field will be best served if those standards place the participant at the center, such that trust can be established by providing clear understanding, choice and control.

6. POTENTIAL APPLICATION FOR PERSONAL SENSING

6.1 Integration into existing models of care

A personal mental health sensing platform with sufficient accuracy could enhance mental health care by helping identify people in need of treatment, accelerating access to treatment, and monitoring functioning during or after treatment. The inability to identify patients in need of treatment is a major failure point in our healthcare system. Nearly 60% of all people with a mental health condition receive no treatment in any given year. Our healthcare system relies almost entirely on people with mental health conditions to present themselves for treatment. Thus, accessing care in a timely manner relies primarily on the patient, whose condition may involve a loss of motivation, stigmatization, a sense of hopelessness and helplessness, in some cases impaired judgment, all of which may interfere with help seeking.

While personal mental health sensing holds great promise for monitoring at-risk populations to deliver care more rapidly and effectively, developing accurate algorithms alone will not solve the problem. This will require user-centric approaches to privacy and user control, as well as providing sufficient value to all end users (patients and providers) to promote use. Such systems will also likely present situations for which there are no care guidelines. For example, if a mental health sensing system forecasts that a bipolar patient has a high likelihood of having a relapse in the next two weeks, would clinic staff know what to do (Mayora et al 2013)? Thus, the ability to detect potential mental health problems opens enormous potential to improve access to care, but the solution will require considerable clinical and design research.
beyond the personal sensing described in this paper.

6.2 Behavioral Intervention Technologies

Behavioral intervention technologies for mental health such as websites or mobile apps consist largely of psychoeducational content (text or video) and interactive tools. Sensing capacity for health behaviors such as physical activity or sleep has resulted in apps that are less reliant on patients to log activities, presumably making apps easier to use and therefore more effective. As attractive as that may sound, here too, previous efforts have demonstrated many unknowns beyond personal sensing (Burns et al 2011). While research has discovered a lot about behavioral and environmental factors that contribute to mental illness, it has produced little granular knowledge about the wishes, goals, challenges, and aspirations of people on a moment to moment, hour to hour basis. This information is critical to design the next generation of intervention technologies that fit into the fabric of people’s daily lives. Applications have tended to be designed with a top-down approach, trying to get people to do what we think will help them. But technologies that are adopted and widely used are commonly those that make some aspect of people’s lives easier, helping people do or achieve something they are motivated to do. Success will be more likely if what is sensed and how sensed data is used speaks to the user’s personal goals, integrating treatment aims with making their goals and tasks easier on a daily basis, thereby fitting treatment activities into people’s common patterns and actions.

6.3 Epidemiology

Databases with genomic, epigenetic and other biological data are being integrated with clinical data bases to explore genetic influences on disease. While there is growing recognition that behavior is a critical factor, behavioral data has traditionally been collected using self-report measures, which provides only a periodic subjective snap shot. Personal sensing
platforms can provide a continuous stream of objective data that can be used to explore interactions between behavioral markers, genetic and biological factors, and disorders.

7. SUMMARY AND CONCLUSIONS

A growing cloud of digital exhaust is emitted from our daily activities and actions. Some of these data are produced intentionally, such as through the use of wearables. But much of it is a byproduct of our daily actions captured through our phones, computers, purchasing, and the increasingly sensor enabled objects in our lives. The promise for research as well as clinical care in mental health is enormous. But the challenges are also large and manifold. While the feasibility of personal sensing for mental health has been demonstrated, enormous challenges remain to move from proof-of-concept to tools that are useful in broader populations. The ultimate success of personal sensing in mental health will likely depend on the continued engagement of users, who supply both passively collected data and some measure of active labeling. This, we believe, will require an infrastructure that is a social machine, sufficiently engaging users to prevent obsolescence. Creating trust in these system will require a recognition of the primacy of the user, instantiated by enabling people to understand, control, and own their data. While the tasks are considerable, the potential benefits are also game-changing. The ability to continuously identify behaviors related to mental health has the potential to transform the delivery of care, speeding recognition of people at risk or in need of treatment, and ushering in a new generation of highly personalized, contextualized, dynamic mHealth tools that can listen rather than ask, seamlessly interacting, learning, and growing with users.
SUMMARY POINTS

1. Using sensors in our everyday lives, personal sensing offers the potential to measure human behavior continuously, objectively, and with minimal effort from the user.

2. Translating raw sensor data into knowledge can be achieved using a layered, hierarchical approach in which sensor data are converted into features, and features are combined to estimate behaviors, moods and clinical states.

3. A growing number of studies have found that phone sensor data (e.g. GPS, accelerometry, light, microphone, etc.) can, using machine learning, provide markers of sleep (e.g. bed/waketime, duration), social context (e.g. who is in vicinity, relationship to in-phone contacts), mood, and stress.

4. Depression and mood states in bipolar disorder have been estimated using variety of phone sensor data. GPS features measuring entropy and circadian rhythm of movement have been correlated with depression.

5. Social media (Facebook, Twitter) posts can identify people who are depressed, or likely to become depressed.

6. While the work on phone sensor data has been promising, most studies have been small, on the order of 7-30 participants, who are frequently college students, with little evidence to support replicability.

7. Machine learning methods vary, some relying on user generated labels while other uncover patterns in unlabeled data. Labeling often improves and helps algorithms adapt to new circumstances. Thus, rather than an autonomous prediction machine, it may be more useful to think of a mental health sensing platform as a social machine in which the quality of prediction is a shared endeavor.
8. Research to date suggests that personal sensing using everyday sensors is feasible.

However, numerous challenges must be overcome before it is viable for clinical deployment.
FUTURE ISSUES

1. Because of the amount of variability coming from differences in hardware, device usage patterns, lifestyle, and environment, personal sensing platforms will likely require a very large user base to have widespread applicability.

2. Some data, such as GPS, is impossible to de-identify while retaining its utility. Thus, creating trust in these system will require a recognition of the primacy of the user, instantiated by enabling people to understand, control, and own their data.

3. Personal sensing offers the potential of a new class of intervention technologies that can reduce user burden while creating highly tailored and contextualized interactions.

4. No sensing system will be 100% accurate and thus researchers, developers, and users must come to consensus about how much error is acceptable and how to better explain and display error to relevant stakeholders.

5. Improving systems will likely require some action on the part of users and ways to ensure that actions are directly associated with benefits will likely create more engaging and empowering systems.

6. Personal sensing can help improve screening and access for treatment but requires advances in infrastructure and integration into workflow as well as improvements in underlying technology, knowledge and improving algorithmic accuracy.

7. The field of personal sensing will likely continue to experience a tension between what is possible and what is feasible related to a trade-off that occurs between small proof-of-concept studies demonstrating novelty and large studies demonstrating robustness and generalizability.

8. Integrating personal sensing data with clinical and genomic databases offers the
opportunity to deepen our understanding of the relationship behavior and genomic X behavior interactions on health, wellness, and disease.
ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Mental Health with grants P20MH090318, R01MH100482, and R01MH095753 to Dr. Mohr, and K08MH102336 to Dr. Schueller.
References


Ben-Zeev D. *UbiCompUnder Review.*


Bayesian Statistics 8: 3-23

Brush AJB, Krumm J, Scott C. *ACM International Conference on Ubiquitous Computing (UbiComp), Copenhagen, Denmark, 2010*. ACM.


Canzian L, Musolesi M. *UniComp '15, Osaka, Japan, 2015*. ACM.


Ciman M, Wac K, Gaggi O. *Pervasive Computing Technologies for Healthcare (PervasiveHealth), Istanbul, Turkey, 2015*. IEEE.


De Choudhury M, Gamon M, Counts S, Horvitz E. *International Conference on Web and Social Media, Boston, MA USA, 2013b*. AAAI.

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. *Computer Human
Interaction, San Jose, CA, USA, 2016. ACM.


Eadicicco L. 2015. Americans check their phones 8 billion times a day. In Time


Lim BY, Dey AK. International Conference on Ubiquitous Computing (UbiComp), Beijing, China, 2011. ACM.


Madan A, Cebrian M, Lazer D, Pentland A. International Conference on Ubiquitous Computing (UbiComp), Copenhagen, Denmark, 2010: 291-300. ACM.


Ricker T. 2015. First Click: Wearables are booming but is anyone still wearing them? *The Verge*


Sen S, Subbaraju V, Misra A, Balan RK, Lee Y. *International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), St. Louis, MO, USA, 2015*: 585-90. IEEE.


Shilton K, Sayles S. *International Conference on System Sciences, Honolulu, HI, USA, 2016.* IEEE.


Sweeney L. 2000. Simple Demographics often identify people uniquely, Carnegie Mellon University, Pittsburgh, PA


Vallance JK, Winkler EA, Gardiner PA, Healy GN, Lynch BM, Owen N. 2011. Associations of


Zhenyu C, Lane N, Cardone G, Lin M, Choudhury T, Campbell A. *Proceedings of Pervasive*

Figure 1. Overarching Framework
ACRONYMS AND DEFINITIONS

Behavioral (Personal?) Sensing: collection and analysis of data from sensors embedded in the context of daily life, with the aim of identifying human behaviors, thoughts, feelings, and traits.

Behavioral Marker: Behaviors, thoughts, feelings, traits, or states identified using personal sensing.

Ubiquitous Sensing: Use of networked sensors weaved into everyday life to capture information about humans, environments, and their interactions anytime and everywhere.

Machine Learning: A subdiscipline of artificial intelligence that builds algorithms that have the ability to learn without explicitly programmed instructions.

Feature: A measurable property of a phenomenon, which is proximal to and constructed from available sensor data.

Label: in machine learning, a label refers to that which is being predicted, similar to a dependent variable in statistics.

GPS: Acronym for global positioning service, although today it sometimes refers location
services that fuse GPS with other signals, such as WiFi, to obtain greater accuracy with less battery drain.

**Curse of Dimensionality**: As the number of dimensions expands, the available data in the space become sparse, which prevents machine learning methods from being efficient.

**Wearable**: A computing technology designed to be worn, many contain embedded sensors for specific purposes, most commonly for health or fitness.

**Sensor**: a device the measures a physical property and produces a corresponding output.

**Social media**: technological tools that allow people, companies, and organizations to share user-generated information and connect with other users through networks.

**Supervised learning**: A category of machine learning that uses labeled data provided by a set of training samples in order to construct an algorithm,

**Unsupervised learning**: A category of machine learning that attempts to uncover underlying structure in the data and does not require labeled data.

**Semi-supervised learning**: A category of machine learning that combines aspects of supervised and unsupervised methods by using samples with both labeled and unlabeled data.
Active Learning: A subset of semi-supervised learning in which the algorithm can query a user to provide additional labels when the algorithm is uncertain how to classify a set of data.