MoTurk: A mobile personal data market for research use

Yining Cao
University of Michigan
Ann Arbor, MI
rimacyn@umich.edu

ABSTRACT

In this paper, we propose a mobile crowd sourcing platform, MoTurk, for transactions of mobile sourced personal data between the general public and researchers through a Mechanical Turk mechanism. We mainly reconcile the dilemma of flexible acquisition of mobile sensing data and privacy protection by summarizing features that support a four layer hierarchical model and implementing a functional programming model for data processing.

KEYWORDS

Personal Data, Privacy, Mobile Sensing

1 INTRODUCTION

Two-thirds of U.S adults own smart phones equipped with various sensors[1]. On the other hand, indications of behavior inferred from mobile sourced data (e.g., sleep[2], conversation[3]) have seen an increasing use in research areas [4]areas. Namely, sleeping hours are strongly related to clinic health and academic performance [2, 5–7], a thorough analysis of location data helps understand mobility pattern of the general public, contributing to urban planning [8, 9] or serving as indications of individuals' life pattern. However, current implementation of mobile sensing data for research use encounters two major difficulties:

- **Data acquisition difficulty:** Researchers typically need to develop a data collector app for collecting mobile sourced data(e.g.,coordinate, audio data). Moreover, as raw sensing data can hardly be used as behavior indications, further data transformation and processing methods are required.
- **Privacy concerns:** Current 'Informed Consent' policy provides participants with less power of privacy control over how and to what extent their personal data are used. There also lack references for tangible value of personal data, making it hard for the public to decide whether or not to share their privacy.

With current efforts of personal data sensing, collecting and sharing, participants over pose their privacy with raw data while researchers take time and efforts to 'degrade' raw data into behavioral indicators or other insightful statistics with a coarser level of data granularity. In hope of filling this gap, this paper presents a platform, MoTurk, as a personal data market. We use Mechanical Turk mechanism to promote transactions of mobile sourced data between the general public and researchers in various research domains.

MoTurk attempts to reconcile the dilemma between data acquisition and privacy protection via the following features:

Online customization : Support data acquisition at different granularity with an expressive data processing scheme;

Feature extractor : Provide various features indicating human behaviors and use in-build feature extractors to transform raw data into these features;

Privacy enforcement : Upload data at defined granularity and relate privacy sensitivity to price through market.

The rest of the paper is organized as follows: Section 2 overviews the related work, Section 3 presents the MoTurk system architecture and system details designed to support major features. Sections 4 concludes the paper by discussing current system and future work.

2 RELATED WORK

To help simplify an App developing process, there are open and well-documented API providers (e.g.Anyplace[10] , OpenStreetMap(OSM)[11]), frameworks(e.g., Funf[12], AWARE[13]) and tool-kits(e.g.,ResearchKit[14]), however, configurations and deployments of native applications are still heavy burdens for non-computer background researchers. Moreover, difficulty of data acquisition also lies in the range of user base, which calls for the implementation of cloud-based platform for collecting and leveraging crowdsourced data. Current online mobile data providers like CrowdSignal.io[15] and CrODA-gator[16] are all based on existing data set rather than a study based test bed. PhoneLab[17] and LabintheWild[18] deal with data for each study by leveraging a research setting stage, yet they provide services for mobile application deployment testing and online experiments respectively, rather than mobile sensing data related to human behaviors. Table1 summarizes key high features of the various previous works and compares them with the proposed MoTurk.

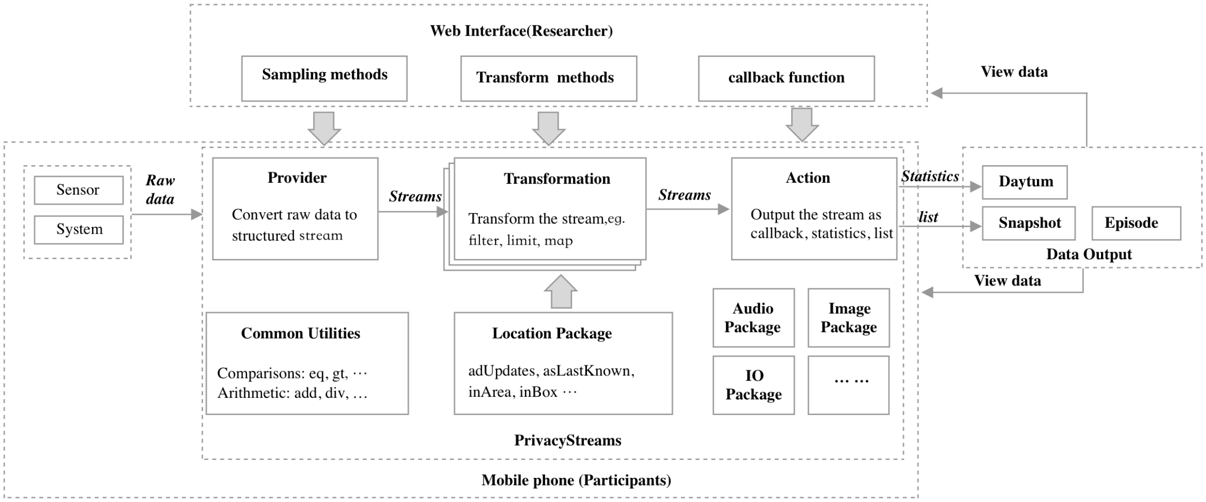
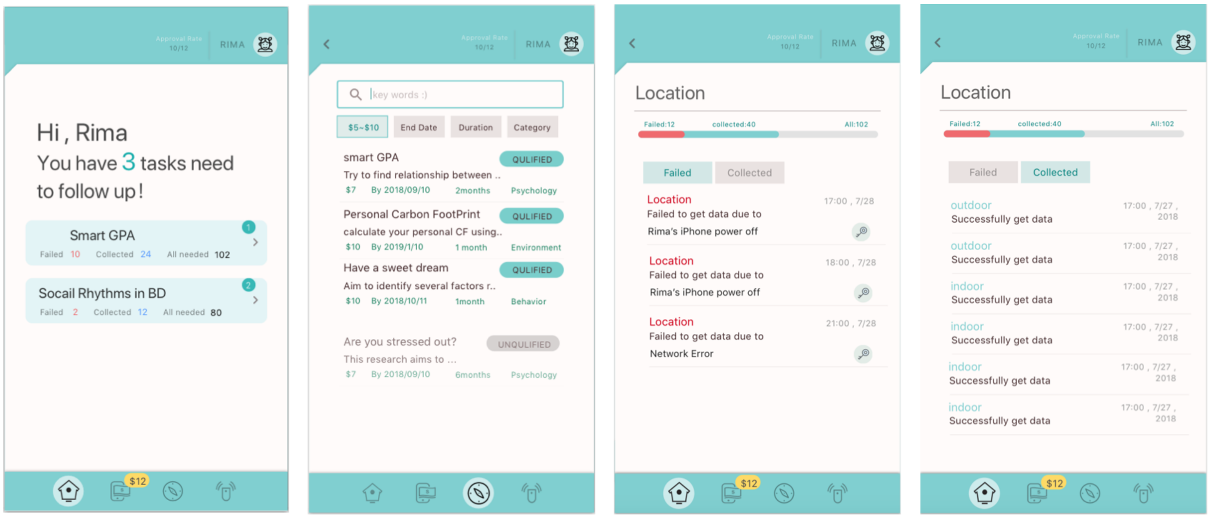


Figure 1: Data processing within PrivacyStreams



Hi , Rima
You have 3 tasks need to follow up !

Smart GPA 1
Failed 10 Collected 24 All needed 102

Socail Rhythms in BD 2
Failed 2 Collected 12 All needed 80

key words :)

$5~$10 | End Date | Duration | Category

smart GPA QULIFIED
Try to find relationship between ..
$7 By 2018/09/10 2months Psychology

Personal Carbon FootPrint QULIFIED
calculate your personal CF using..
$10 By 2019/1/10 1 month Environment

Have a sweet dream QULIFIED
Aim to identify several factors r..
$10 By 2018/10/11 1month Behavior

Are you stressed out? UNQULIFIED
This research aims to ...
$7 By 2018/09/10 6months Psychology

Location
Failed:12 collected:40 All:102
Failed | Collected

Location 17:00 , 7/28
Failed to get data due to
Rima's iPhone power off

Location 18:00 , 7/28
Failed to get data due to
Rima's iPhone power off

Location 21:00 , 7/28
Failed to get data due to
Network Error

Location
Failed:12 collected:40 All:102
Failed | Collected

outdoor 17:00 , 7/27 , 2018
Successfully get data

outdoor 17:00 , 7/27 , 2018
Successfully get data

indoor 17:00 , 7/27 , 2018
Successfully get data

indoor 17:00 , 7/27 , 2018
Successfully get data

indoor 17:00 , 7/27 , 2018
Successfully get data

indoor 17:00 , 7/27 , 2018
Successfully get data

Figure 2: The MoTurk Application Screen shots for study viewing, study notification and data viewing

On privacy, there are works for checking and maintaining privacy policies, like Sensibility Testbed[19], OpenPDS[20]. While Sensitive Testbed simply avoid sensitive data or encrypt by hashing to alleviate privacy sensitivity, OpenPDS protects users' privacy by offering them a paradigm called SafeAnswers, allowing customization before data sharing. However, they realize it by allowing researchers and applications to submit code (the question) to be run against the metadata. The full accessibility may thus pose threat to privacy.

There are several behavioral researches have already been done on Amazon Mechanical Turk platform [21–23], supporting it as an ideal labor market for experiments, seldom do they use the mobile sensing data. In terms of financial compensations, prices vary from $5 per week [24] to $50 per week [25] without references. Though researches have done and some conclusions have drawn in this area(e.g., people value location data most [26]), tangible values of personal data remain controversial due to a confined lab environment and small sample size.

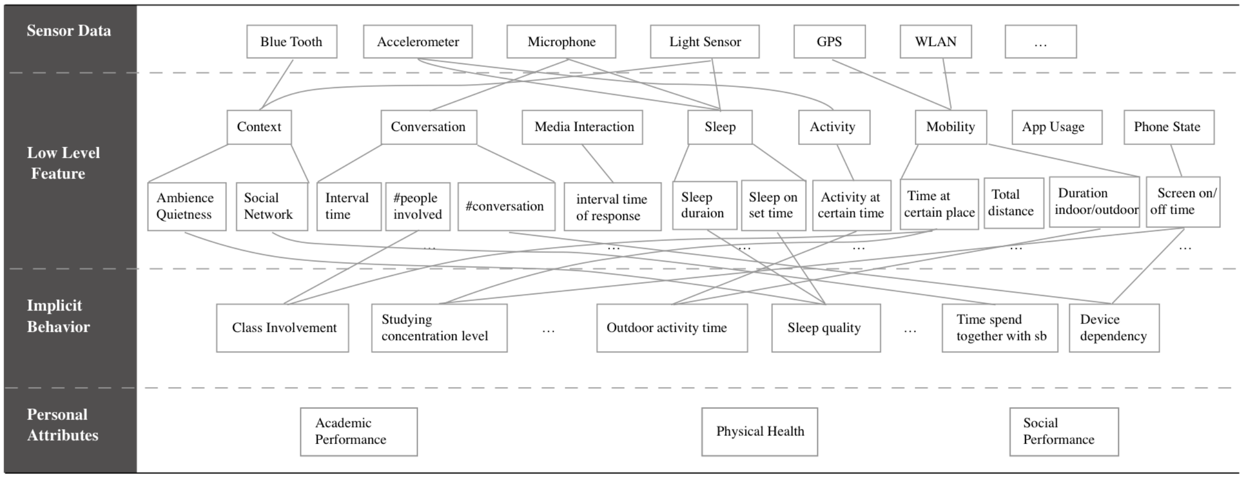
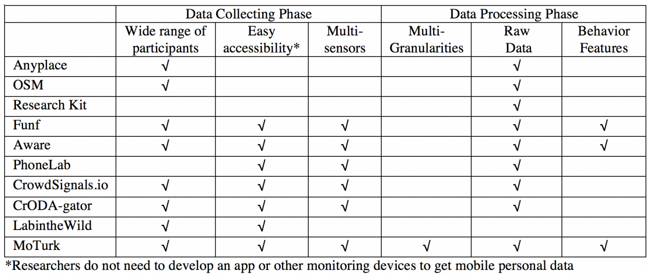


Figure 3: Data transformation flow of understanding students' performance from data features

Table 1: Comprehensive comparisons between MoTurk and previous platforms



	Data Collecting Phase			Data Processing Phase		
	Wide range of participants	Easy accessibility*	Multi-sensors	Multi-Granularities	Raw Data	Behavior Features
Anyplace	√				√	
OSM	√				√	
Research Kit					√	
Funf	√	√	√		√	√
Aware	√	√	√		√	√
PhoneLab		√	√		√	
CrowdSignals.io	√	√	√		√	
CrODA-gator	√	√	√		√	
LabintheWild	√	√				
MoTurk	√	√	√	√	√	√

*Researchers do not need to develop an app or other monitoring devices to get mobile personal data

3 PROPOSED SOLUTION

3.1 System Architecture

This section outlines the MoTurk work-flow.

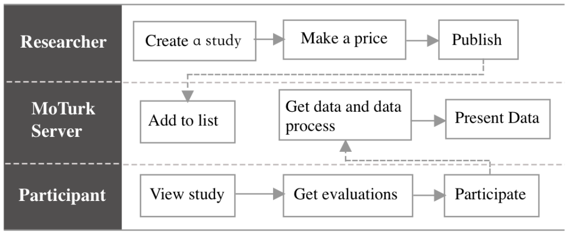


Figure 4: System architecture for MoTurk

As shown in Figure 4, MoTurk contains three stakeholders, researchers, users and MoTurk server respectively

In the*study setting stage*, researchers will be able to create studies with specific research needs including types of data(mobility data, conversation data,etc.), requests for participants(age, occupation, etc.) and expected data granularity, sampling and data processing methods. Evaluations concerning privacy sensitivity and battery impact are also generated in this stage as a reference for price making.

After a study has published, it will be pushed to the server and users who are qualified for this study are able to see it through MoTurk App on mobile phones. When viewing interested studies, users will be fully access to the information researchers created including basic information, required personal data, price and evaluations. When they finally choose to start this study, the *data collecting stage* automatically begins. their data begin to be collected and processed . All data will be updated and presented to both researchers and users with a 24 hour interval. Users will be paid at the end of studies.

Interfaces for researchers and users are separate, while researchers do all operations on the web interface, users interacts with mobile phone with MoTurk App.

3.2 Contributions and Details

In this section we will highlight two major contributions of our system followed by details.

3.2.1 Hierarchical data . To summarize meaningful behavioral features and statistics of mobile sourced data for research use, we covered 81 research papers in 6 domains. As is shown in Table 2Data in need vary from different research domains of and mostly used kinds of data fall into 7 categories: Context, Conversation, Media Interaction, Sleep, Mobility, App Usage, Phone Usage. Our platform attempts to extract and provide most useful features for research needs in each of these 7 categories.

Table 2: Summarize of mobile sourced data usage in each discipline for human behavior indications

Discipline[a]	Topics[b]	Mostly used type of data
Behavioral Science (23)	Relationship [27];Life pattern [28];Well being [5]	Mobility, Activity, Conversation
Psychology (15)	Personality [29];Mental disease detection [25]	Conversation, Sleep, Phone State
Environmental Science (14)	Environmental impact evaluation [30]; Urban traffic planning [9]	Mobility, Activity, Context
Cognitive Science (13)	Detection of Alertness [24],Stress [31],Boredom [32],Attention [33], etc.	App Usage, Context
Marketing (9)	Advertising strategy [34]; Location-based service [35]	Mobility, Context
Software Development (7)	User behavior [36]; Visual reality [37]	Context, App usage

[a] number in parentheses represents number of research paper covered in certain discipline.

[b] Restricted by space, we list one of many papers as reference for each topic in the table.

We stratify mobile sourced personal data into a four layer hierarchical model [27], ranging from raw data to behavioral indications:

Raw data : data sampled directly from mobile sensors like accelerators, microphone, GPS, etc.

Low Level Features : meaningful statistics drawn from raw data which can represent or represent specific human behaviors;

Implicit Behaviors : inferred daily behaviors of human beings within a specific scenario. We draw a line between the second and third level simply by that the third level of data is the combination of several low level features.

Personal Implications : this level of data is not simply statistics but implications, which can be rather abstract and refer to attributes of users themselves.

MoTurk support both row data and low level features according to our previous discussion, enabling researchers to get exactly kinds of data they need. Here we take an example scenario to explain the capability of various data for research needs we support. Figure 3 shows a data transformation flow of understanding students' performance from data features.

for a clear presentation, all lines of relationship are not drawn completely, and no lines were drawn between the fourth level and the third level, indicating that personal attributes may not be merely inferred by implicit behaviors but can also be inferred by low level features or even raw data [28, 29].

3.2.2 Customization with privacy enforcement. To simplify both the data collection and processing process, MoTurk adopts an expressive personal data processing scheme, PrivacyStreams[30], to allow for rich customized needs by specifying parameters. This guarantees researchers can get the data in need without too much further data processing. Basic work flow is as shown in figure1

The basic form for getting data through PrivacyStreams:

```
UQI.getData(Provider)
   .data transform method1()
   .data transform method2()
      ... ...
    .action(callback)
```

A more concrete scenario of accessing raw audio every two minutes for 10 seconds and transforming it into loudness is as follows:

```
UQI.getData(Audio.recordPeriodic(10*1000,2*60*1000))
    .setField("loudness",
     calcLoudness(Audio.AUDIO_DATA))
    .onChange("loudness", callback)
```

PrivacyStreams provides a scalable and customizable structure for data processing. The stream processing model of PrivacyStreams allows customization of transformation methods and callback functions, also makes it possible to analyze how personal data is processed and what granularity of data is actually used. We enforce privacy by processing data transparently and locally.

4 DISCUSSION

- **Win-Win solution:** Identified personal data behavior features and an expressive data processing scheme allow researchers to collect different types of data in defined granularity while ensuring transparency and privacy.
- **Platform as a Service:** As a Crowdsourcing platform, MoTurk reinforces interactions between researchers and participants by keeping them informed of why and how their sensitive resources are used and thus maintains a stable user base. However, as MoTurk needs a large user base to make the transaction possible, strategies are needed for developing a large test bed.
- **Evaluation model:** More research works are needed for develop models for privacy sensitivity and battery impact of a certain study with configurations set by researchers. Besides, annoying rate of a study also contribute to financial compensations;

REFERENCES

[1] Kelly Servick. Mind the phone, 2015.

[2] Steven P Gilbert and Cameron C Weaver. Sleep quality and academic performance in university students: A wake-up call for college psychologists. *Journal of College Student Psychotherapy*, 24(4):295–306, 2010.

[3] Danny Wyatt, Tanzeem Choudhury, Jeff Bilmes, and James A Kitts. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):7, 2011.

[4] Gabriella M Harari, Sandrine R Müller, Min SH Aung, and Peter J Rentfrow. Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, 18:83–90, 2017.

[5] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 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*, pages 3–14. ACM, 2014.

[6] Daniel J Taylor, Karlyn E Vatthauer, Adam D Bramoweth, Camilo Ruggero, and Brandy Roane. The role of sleep in predicting college academic performance: Is it a unique predictor? *Behavioral sleep medicine*, 11(3):159–172, 2013.

[7] Adrienne Wald, Peter A Muennig, Kathleen A O'Connell, and Carol Ewing Garber. Associations between healthy lifestyle behaviors and academic performance in us undergraduates: a secondary analysis of the american college health association's national college health assessment ii. *American Journal of Health Promotion*, 28(5):298–305, 2014.

[8] Jing Dong, Changzheng Liu, and Zhenhong Lin. Charging infrastructure planning for promoting battery electric vehicles: An activity-based approach using multiday travel data. *Transportation Research Part C: Emerging Technologies*, 38:44–55, 2014.

[9] Colin Sheppard, Rashid Waraich, Andrew Campbell, Alexei Pozdnukhov, and Anand Gopal. Modeling plug-in electric vehicle charging demand with beam, 2017.

[10] Kyriakos Georgiou, Timotheos Constambeys, Christos Laoudias, Lambros Petrou, Georgios Chatzimilioudis, and Demetrios Zeinalipour-Yazti. Anyplace: A crowdsourced indoor information service. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 291–294. IEEE, 2015.

[11] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *Ieee Pervas Comput*, 7(4):12–18, 2008.

[12] N Aharony, Alan Gardner, C Sumter, and A Pentland. Funf: Open sensing framework, 2011.

[13] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[14] Jennifer Jardine, Jonathan Fisher, and Benjamin Carrick. AppleâĂŹs researchkit: smart data collection for the smartphone era?, 2015.

[15] Evan Welbourne and Emmanuel Munguia Tapia. Crowdsignals: a call to crowdfund the community's largest mobile dataset. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 873–877. ACM, 2014.

[16] Michalis Massalas, Andreas Konstantinidis, Achilleas Achilleos, Christos Markides, and George Papadopoulos. Croda-gator: An open access crowdsourcing platform as a service.

[17] Anandatirtha Nandugudi, Anudipa Maiti, Taeyeon Ki, Fatih Bulut, Murat Demirbas, Tevfik Kosar, Chunming Qiao, Steven Y Ko, and Geoffrey Challen. Phonelab: A large programmable smartphone testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining*, pages 1–6. ACM, 2013.

[18] Katharina Reinecke and Krzysztof Z Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1364–1378. ACM, 2015.

[19] Yanyan Zhuang, Albert Rafetseder, Yu Hu, Yuan Tian, and Justin Cappos. Sensibility testbed: Automated irb policy enforcement in mobile research apps. In *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications*, pages 113–118. ACM, 2018.

[20] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S Wang, and Alex Sandy Pentland. openpds: Protecting the privacy of metadata through safeanswers. *PloS one*, 9(7):e98790, 2014.

[21] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. 2010.

[22] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[23] Winter Mason and Siddharth Suri. Conducting behavioral research on amazonâĂŹs mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[24] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 178–189. ACM, 2016.

[25] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association*, 23(3):538–543, 2016.

[26] Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo de Oliveira, Michele Caraviello, and Nicu Sebe. Money walks: a human-centric study on the economics of personal mobile data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 583–594. ACM, 2014.

[27] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[28] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 351–360. ACM, 2012.

[29] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):5, 2017.

[30] Yuanchun Li, Fanglin Chen, Toby Jia-Jun Li, Yao Guo, Gang Huang, Matthew Fredrikson, Yuvraj Agarwal, and Jason I Hong. Privacystreams: Enabling transparency in personal data processing for mobile apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):76, 2017.