CCF Transactions on Pervasive Computing and Interaction (2021) 3:344–366
https://doi.org/10.1007/s42486-021-00072-4

REGULAR PAPER

Understanding practices and needs of researchers in human state modeling by passive mobile sensing

Xuhai Xu[1] · **Jennifer Mankoff**[1] · **Anind K. Dey**[1]

Received: 5 January 2021 / Accepted: 4 June 2021 / Published online: 6 July 2021

Abstract
Passive mobile sensing for the purpose of human state modeling is a fast-growing area. It has been applied to solve a wide range of behavior-related problems, including physical and mental health monitoring, affective computing, activity recognition, routine modeling, etc. However, in spite of the emerging literature that has investigated a wide range of application scenarios, there is little work focusing on the lessons learned by researchers, and on guidance for researchers to this approach. How do researchers conduct these types of research studies? Is there any established common practice when applying mobile sensing across different application areas? What are the pain points and needs that they frequently encounter? Answering these questions is an important step in the maturing of this growing sub-field of ubiquitous computing, and can benefit a wide range of audiences. It can serve to educate researchers who have growing interests in this area but have little to no previous experience. Intermediate researchers may also find the results interesting and helpful for reference to improve their skills. Moreover, it can further shed light on the design guidelines for a future toolkit that could facilitate research processes being used. In this paper, we fill this gap and answer these questions by conducting semi-structured interviews with ten experienced researchers from four countries to understand their practices and pain points when conducting their research. Our results reveal a common pipeline that researchers have adopted, and identify major challenges that do not appear in published work but that researchers often encounter. Based on the results of our interviews, we discuss practical suggestions for novice researchers and high-level design principles for a toolkit that can accelerate passive mobile sensing research.

Keywords Researchers · Practices and needs · Human state modeling · Mobile sensing

1 Introduction

Understanding and modeling human behavior is essential for many studies in health, social science, and behavior sciences (Harari et al. 2017). Because of the difficulties of continuously collecting longitudinal behavioral data in the wild (Reis 2012), previous studies mainly investigated human daily behavior in the real world using an ethnographic methodology before the emergence of smartphones. Nowadays, smartphones, embedded with a number of rich sensors, have become ubiquitous and serve as a powerful sensing platform for passively capturing human behavior (Lane et al. 2010) at an unprecedented resolution.

In the past decade, researchers have been investigating how passive mobile sensing can be used for human state modeling, by applying this technique to many longitudinal behavior-related problems. It's been applied to a wide range of fields: physical and mental health monitoring (Carreiro et al. 2015; Canzian and Musolesi 2015), affective computing (Zhang et al. 2018; Suhara et al. 2017; Mehrotra et al. 2017), academic/work performance evaluation (Mirjafari et al. 2019; Wang et al. 2015), routine modeling (Banovic et al. 2016; Qin et al. 2018), etc. The main focus of previous research has been on demonstrating the value of the passive sensing technique. A typical framing of a research question is: whether passive mobile sensing can model/detect/predict X, where X belongs to a kind of human state, such as emotions, mental health status, or substance usage. However, much less focus has been placed on the researchers

✉ Anind K. Dey
anind@uw.edu

Xuhai Xu
xuhaixu@uw.edu

Jennifer Mankoff
jmankoff@cs.uw.edu

[1] Information School, University of Washington, Seattle, WA 98105, USA

themselves. Researchers themselves do not often appear in their publications with their detailed practices, challenges, and needs often hidden beneath their papers. Actually, many interesting and informative questions can be asked about researchers. For example, what are the methodologies they followed when conducting passive sensing research? What are the common problems and difficulties they encountered when running their study? What are the needs they had when collecting and analyzing the data? Existing publications do include a research procedure, such as how authors collected and analyzed the data, which indicates researchers' practices to some extent. However, many details are not included in papers, especially the difficulties and obstacles that do not directly contribute to the research goals but have to be solved by researchers nonetheless. For example, a typical publication about the application of passive mobile sensing often briefly describes the data collection procedure with a few paragraphs (e.g., (Bae et al. 2017; Xu et al. 2019)). It does not introduce the challenges the authors encountered during the data collection study and how they solved them (e.g., the difficulties of application development and data quality monitoring), because these are usually not the focus of their paper. However, knowing the challenges and potential solutions can help researchers who are new to this area, allowing them to conduct similar studies without so much pain.

In this paper, we address this gap by conducting interviews with experienced researchers in this area to deeply investigate their practices, challenges, and needs. Passive mobile sensing research is attracting more attention, and a larger number of researchers outside this area are looking to apply this approach to their problem domains. Answering these often hidden questions about researchers is of great importance and can benefit a wide range of audiences, especially for *researchers who are interested in this research area but do not have much previous experience*. First, our paper provides an overview of the current research practices and existing obstacles in the field of passive mobile sensing. There exists a few review papers on passive mobile sensing techniques and their applications (Choudhury et al. 2008; Harari et al. 2017; Trifan et al. 2019; Laport-López et al. 2020; Lane et al. 2010) but they do not cover researchers' common practices and challenges. Our paper complements these review papers and provides valuable lessons for laymen and beginning researchers to better understand the state-of-the-art. Moreover, intermediate researchers can also benefit from the answers. They can learn from others to polish their own study pipelines, avoid unnecessary pitfalls, and better hone in on important research questions. Finally, an understanding of the practices and needs can further inform the design of a toolkit for assisting a wide range of researchers who plan to apply passive sensing in their studies.

Current passive mobile sensing research requires a significant amount of technological knowledge. Although researchers from psychology, social science, behavior science are often involved as collaborators, the majority of the researchers in this area have a technology background. Therefore, in this interview, our scope is focused on technical researchers, in the fields of computer science, electrical engineering, and information science. We interviewed ten experienced researchers that have at least 2 years of research experience and have at least one first-author publication in the past 3 years in top venues (including IMWUT, WWW, and JMIR).

Our interview results reveal a number of interesting details about researcher practices and pain points encountered by researchers that are not often discussed in their publications. For instance, the majority of the participants (seven out of ten) collected their own datasets instead of using a public dataset, during which a dashboard was commonly used to monitor data collection. Mobile application development, debugging hardware/software issues, communication with study participants, and workload were major challenges that stood out during the interview. Researchers also struggled when conducting data cleaning and data analysis, due to difficulties of data heterogeneity, model selection, vague evaluation, communication among the team, etc. Moreover, participants further expressed a wide range of needs, such as missing data imputation, outlier detection, feature extraction, and algorithm selection, which can potentially be assisted by a new toolkit. Their challenges and needs shed light on the design guidelines of such a toolkit.

Our paper is structured as follows. Section 2 provides a brief background of passive mobile sensing for human state modeling. Section 3 introduces the interview protocol design, participants, the procedure, and the analysis method. Section 4 summarizes the common practices conducted by researchers during their research. Laymen and beginning researchers can best benefit from these results to learn about the state-of-the-art. Section 5 summarizes the common pain points and needs researchers often encounter. Intermediate researchers may find these results most meaningful for them to improve their own studies and research pipelines. Section 6 discusses important roles that researchers play in the research process and addressing the uniqueness of passive mobile sensing data. We envision that researchers from any background can potentially find these results interesting, provoking more reflections on this area. Section 7 discusses the implications for practical suggestions, toolkit design guidelines, as well as the limitations and future work. Finally, Sect. 8 concludes the paper.

2 Background

Nowadays, people carry their smartphones and wearables almost every day to almost everywhere they go. This equips these devices with the ability to passively capture,

understand, and model people's daily behaviors at an unprecedented resolution (Lane et al. 2010). Researchers have applied a passive mobile sensing technique to a wide range of areas in human state modeling, such as health monitoring (Carreiro et al. 2015; Canzian and Musolesi 2015; Biel et al. 2018; Chang et al. 2018), affective computing (Zhang et al. 2018; Suhara et al. 2017; Mehrotra et al. 2017), activity recognition (Yan et al. 2012; Sun et al. 2017; Vaizman et al. 2017), crime detection (Bogomolov et al. 2014, 2015), financial behavior prediction (Centellegher et al. 2018; Di Clemente et al. 2017), academic/work performance evaluation (Mirjafari et al. 2019; Wang et al. 2015), and routine modeling (Banovic et al. 2016; Qin et al. 2018), to name just a few.

As the adoption of passive sensing techniques becomes more widespread, the amount of literature has been growing quickly even within one single application domains. Take the area of healthcare for example. Recently, researchers have leveraged passive sensing for both physical and mental health (Madan et al. 2012; Lane et al. 2010). Example topics from the broad category of physical health include daily activities, such as sleep (Min et al. 2014; Chang et al. 2018; Sun et al. 2017), smoking (Shoaib et al. 2015; Naughton et al. 2016), cocaine usage (Carreiro et al. 2015, 2016; Chintha et al. 2018), and drinking (Bae et al. 2017, 2018), as well as physical diseases, such as Parkinson's disease (Mazilu et al. 2016; Postolache and Postolache 2019), diabetes (Alexander et al. 2017; Sarda et al. 2019), and even on COVID-19 (Cho et al. 2020; Oliver et al. 2020). As for mental health, examples include emotion recognition (Zhang et al. 2018; Mehrotra et al. 2017), depression detection (Canzian and Musolesi 2015; Salekin et al. 2018; Lu et al. 2018; Xu et al. 2019), schizophrenia diagnosis (Wang et al. 2016b, 2017; Ben-Zeev et al. 2016), and so on. However, despite the large diversity of research topics, mobile sensing data and analysis methods share many similarities among the different studies. We illustrate this by showing two examples in detail.

Wang et al. (2014) collected mobile sensing data from 48 undergraduate and graduate students over 10 weeks to investigate how mobile sensor data could reveal students' life experiences. They collected data from a number of sensors: accelerometer, GPS, WiFi, Bluetooth, microphone, light, and phone screen. They also inferred higher-level behavior data from basic sensor streams: activity (based on accelerometer and GPS), conversation (base on sound), and sleep (a combination of accelerometer, light, sound, and screen). In addition, they used self-reported surveys to collect students' academic records and mental health conditions. After the data collection, they extracted a number of behavior features from the data, and conducted a correlation analysis between the features and students' mental health survey scores. They identified a number of significant correlations, e.g., conversation frequency is negatively correlated with stress level. In the second example, Bae et al. (2017) employed the passive mobile sensing technique to detect alcohol consumption. They collected data from 38 young adults over 28 days. The sensors included the accelerometer, gyroscope, phone screen, on-screen keyboard, battery, call and message, light, network traffic, GPS, proximity sensor, telephony, microphone, mobile application, WiFi, and Bluetooth, all from participants' own mobile phones. Similar to the first example, they also inferred higher-level behavior data of physical activity and conversation. They collected self-reported drinking episodes and the number of drinks consumed from participants as ground-truth. After extracting features from the data, they conducted a correlation analysis between the features and the amount of alcohol that young adults drank, and trained machine learning classifiers to detect drinking episodes.

Although these two examples focused on completely different topics, many sensors and features overlapped. Both studies leveraged data from the accelerometer, GPS, WiFi, Bluetooth, microphone, and phone screen. This overlap is fairly common across passive sensing studies involving mobile phones as most mobile phones have relatively uniform types of sensors. We further conducted a small-scale survey to identify some of these common sensors and features used in studies, as summarized in Table 1. This indicates that there is a lot of similarities in data collection and feature extraction across different mobile sensing studies. We refer readers to Harari et al. (2017), Trifan et al. (2019), and Laport-López et al. (2020)for more complete surveys. In addition to similarities in the data being collected, researchers also leverage common techniques such as correlation analysis (e.g., Wang et al. 2014, 2015; Bae et al. 2017; Zhang et al. 2018; Mehrotra et al. 2017; Wang et al. 2017), and building machine learning models (e.g., Wang et al. 2015; Bae et al. 2017; Doryab et al. 2019a; Wang et al. 2017; Saeb et al. 2015). The similarities and overlap in these publications suggest common practices and challenges that researchers have in their research. However, these are often not directly reflected in their publications, which focus more on the novel application areas, new algorithms, and data insights mined from the data. To our knowledge, our work is the first to summarize the common practices and needs of researchers in this area.

Interviews like the ones we are conducting in the area of passive mobile sensing, have also been conducted in related areas. Back in 2004, Klemmer et al. (2004) interviewed nine tangible user interface (TUI) researchers to inform the design of a toolkit for tangible input. Their results revealed a few challenges that researchers encountered, including the massive amount of programming when deploying novel ubiquitous hardware, the vague association between the interaction design and the specific software implementation, the difficult debugging process to figure out sensing

Table 1 Mobile phones and wearable devices sensors and corresponding features

Sensor	Features	Literature
Mobile app	Frequency and duration of use of individual app/app category, number of changes between app, number of app running	Bae et al. (2017); Zhang et al. (2018); Mehrotra et al. (2017)
Battery	Battery status, charging time, length of charge	Bae et al. (2017)
Bluetooth	Co-location (number of device nearby)	Wang et al. (2014); Bae et al. (2017)
Communication	Incoming/outgoing call duration[#]/count, incoming/outgoing message length[#]/count, message sent-to-receive ratio, number of contacts	Bae et al. (2017); Zhang et al. (2018); Mehrotra et al. (2017); Wang et al. (2017)
GPS	Travel distance, number of cluster, number of place visited, location entropy, normalized entropy, duration in a certain location, number of changes in location, radius of gyration, circadian movement, transition time	Wang et al. (2014); Bae et al. (2017); Zhang et al. (2018); Mehrotra et al. (2017); Wang et al. (2017); Saeb et al. (2015)
Keyboard	Speed of typing, number of insert/delete, number/types of emojies, frequent time slots of typing	Bae et al. (2017)
Light	Brightness[#], dark ratio, bright ratio	Wang et al. (2014); Bae et al. (2017); Zhang et al. (2018); Wang et al. (2017)
Notification	Number of all/accepted notification, seen/decision/response time	Mehrotra et al. (2017)
Proximity	Screen proximities[#]	Bae et al. (2017)
Screen	Number/duration of lock/unlock status, number of single click, long click and scrolls	Wang et al. (2014); Bae et al. (2017); Zhang et al. (2018); Mehrotra et al. (2017); Wang et al. (2017); Saeb et al. (2015)
Sound	Number of conversations, length of conversation, audio amplitude[#], noise-ratio, silence-ratio	Bae et al. (2017); Wang et al. (2014, 2017)
WiFi	Indoor location	Wang et al. (2014); Bae et al. (2017); Zhang et al. (2018); Wang et al. (2017)
IMU[w]	Magnitude/variance of acceleration/angular speed[#], number of steps, number of activity bouts, length of activity bouts[#], number of steps in active bouts	Bae et al. (2017); Zhang et al. (2018); Wang et al. (2017); Doryab et al. (2019a)
Heartrate[w]	Heart rate[#], absolute/negative/positive change[#], number of no change	Doryab et al. (2019a)
Activity[h]	*[accelerometer, GPS, WiFi]* activity type, number of activities, changes in activity[#], indoor activity duration	Wang et al. (2014); Bae et al. (2017); Mehrotra et al. (2017); Wang et al. (2017)
Sleep[h]	*[accelerometer, GPS, WiFi, sound, light, scree]* sleep duration, onset time, wake time	Wang et al. (2014, 2017); Zhang et al. (2018)

[w] Indicates that the sensor is usually collected from wearable devices

[h] Represents high-level sensor that synthesizes information from multiple sensors

[#] Indicates the calculation of min, max, mean, standard deviation, etc

error. In the work of Carter et al. (2008), they interviewed 28 developers in ubiquitous computing. Nine of them worked on mobile systems. The results showed that developers were concerned about valid evaluations and focused on field studies. However, it was difficult for developers to develop robust prototypes in uncontrolled settings, especially when deploying their applications to more than one type of device and across different infrastructures. Meanwhile, researchers also reported the difficulties of gathering data in field experiments. More recently, Min et al. (2016) conducted in-depth interviews with seven experienced mobile developers and conducted an online survey with 46 developers to understand their challenges when developing power-efficient mobile sensing applications. Developers' key challenges were from the significant time and effort for repetitive power measurements since power use needed to be evaluated under a range of real-world usage scenarios and sensing parameters. To our knowledge, there is no previous work specifically focusing on researchers in the area of passive mobile sensing.

3 Interview with researchers

To obtain first-hand information about current research practices and needs, we conducted interviews with experienced researchers in the area of passive mobile sensing for human state modeling. In this section, we describe our inclusion criteria (Sect. 3.1), interview protocol (Sect. 3.2),

Table 2 Demographics of researchers who participated in the interview

ID	Institute category	Country	Gender	Position	Experience (y)
1	University	America	F	Ph.D. student	3
2	University	America	F	Ph.D. student	3
3	University	America	M	Ph.D. student	3
4	University	China	F	Post-doc	7
5	University	America	M	Post-doc	8
6	University	America	F	Assistant Professor	12
7	Company	America	M	Researcher	2
8	University	Finland	M	Researcher	6
9	Company	America	M	Researcher	11
10	Research Institute	Italy	M	Researcher	15

interview procedure (Sect. 3.3), as well as our analysis method (Sect. 3.4).

3.1 Researcher participants

3.1.1 Inclusion criteria

As mobile sensing is a technology-heavy area, almost all research teams involving passive mobile sensing have technical researchers as core members (e.g., (Xu et al. 2019; Doryab et al. 2019a; Wang et al. 2014; Mirjafari et al. 2019)). Therefore, we focus on experienced researchers in this area who have a technical background. We went through the past 3 years' proceedings of three top relevant academic venues: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Proceedings of the Web Conference (WWW), and Proceedings of the Journal of Medical Internet Research (JMIR). We identified close to one hundred papers related to passive mobile sensing, and sent invitation emails to 45 first authors as they had a background in computer science, electrical engineering, information science, or related fields.

3.1.2 Participants

We sent emails to 45 researchers and 10 of them replied with the willingness to participate in an interview (Male = 6, Female = 4, Age = 29.4 $\pm$ 5.9). The ten researchers are from different research teams in four countries. At the time the interviews were conducted, three of them were Ph.D. students, two were postdoctoral researchers, one was an assistant professor and four were senior researchers in research institutions or companies. All participants claimed research expertise in the area of longitudinal human state modeling using passive mobile sensing, and have at least 2 years of research experience in the area (7.0 $\pm$ 4.2 years). Table 2 summarizes the demographic information of participants.

3.2 Interview protocol design

We used a semi-structured interview protocol. The interviewer followed the protocol and asked follow-up questions based on participants' answers. We conducted four iterations of the interview protocol based on two pilot interviews with colleagues. One of the authors first developed an initial protocol, discussed and iterated it with two other co-authors. Then, the authors conducted two rounds of pilot interviews, discussions, and iteration processes, until reaching a consensus on the final protocol. Table 3 lists the questions from the final version.

3.2.1 Project selection

Overall, the protocol consisted of five sections. In the first section, participants were asked to choose one of their recent projects (from the past 3 years) that they were deeply involved in and to provide a high-level overview of the project. The rest of the interview was based on the project they chose.

3.2.2 Detailed procedures

Next, the protocol had three sections about detailed research procedures, in the order of data collection, data cleaning, and data analysis. In these sections, participants were asked to walk the interviewer through these procedures in their selected project, respectively. If participants did not perform one of the steps, that section was skipped. The interviewer asked the initial questions shown in the protocol, e.g., how to deal with missing data and erroneous data in the data cleaning procedure and followed up with deeper questions as needed. For each section, participants were asked an explicit question about the "pain points" they had during each procedure, to make sure that researchers deliberately reflected on the difficulties,

Table 3 Protocol of the semi-structured Interview

Project overview	
1.	Please pick one of your own projects in the past 3 years that you are most familiar with, what are the research goal and research questions in this project?
2.	What is your role and your responsibility in this project?
3.	What is your general approach for solving the research problem?
Data collection	
4.	What is your data source? [If the data is collected by themselves], how do you run the study?
5.	Walk me through your process for getting the data.
6.	What kind of tools, processes, and automated jobs do you use to get data, if any?
7.	During this process, what is your pain points, or bottlenecks that slow you down?
Data cleaning	
8.	Can you walk me through your process for preparing the data?
9.	How did you clean that data, if needed?
10.	[If missing data was mentioned] How do you deal with the missing data?
11.	[If erroneous data was mentioned] How do you deal with the wrong data?
12.	What kind of tools, processes, and automated jobs do you use to prepare data for the analysis, if any?
13.	During this process, what is your pain points, or bottlenecks that slow you down?
Data analysis	
14.	Can you walk me through your process for analyzing the data?
15.	[If feature extraction is mentioned] How do you decide what features to extract in this project? How do you verify whether a feature is useful or effective?
16.	[If feature selection is mentioned] How do you decide the criteria?
17.	How do you determine what analysis methods to use?
18.	What kind of tools, processes, and automated jobs do you use to analyze data, if any?
19.	Given these data, what is your methodology for data analysis?
20.	During your analysis, what is your pain points, or bottlenecks that slow you down?
Pipeline summary	
21.	Through our interview, I noticed that you mentioned {based on the previous answers} bottlenecks. Imagine there is a Tool X can do something for you, what would that be? If you have multiple ideas in mind, you can say it one by one.
22.	During your analysis, which part you think is the essential part that cannot be automated?
23.	What is the unique part of the behavior data, compare to other machine learning projects you have done?

challenges, and needs they encountered during their projects.

3.2.3 Summary

The interview ended with a summary section. First, the interviewer summarized the challenges mentioned by participants in Sect. 3.2.2 to confirm them, and then asked about the assistance that researchers would like from a supportive toolkit. Participants were encouraged to think broadly in these responses and not to place limits on a future tool's abilities. This was designed to further explicitly inquire about the need of participants across their whole project. Finally, the interview closed with two questions: (1) one about the specific parts of the research that participants wanted to do themselves rather than relying on any tool. This helped us to understand researchers' essential roles (i.e., cannot be replaced by an automated tool) across the whole research procedure; (2) one about the differences between passive mobile sensing behavior datasets and other ordinary machine learning datasets. This helped to understand how they viewed the uniqueness of data in this area.

3.3 Procedure

This study was approved by our University IRB. All interviews were conducted online and audio-taped after participants signed the consent form. In the beginning, the interviewer first briefly introduced the purpose of the interview. Then, the interviewer went through the questions listed in the interview protocol and followed up with deeper questions accordingly. Interviews lasted approximately 45 minutes on average. Each participant received a $15 gift card for compensation.

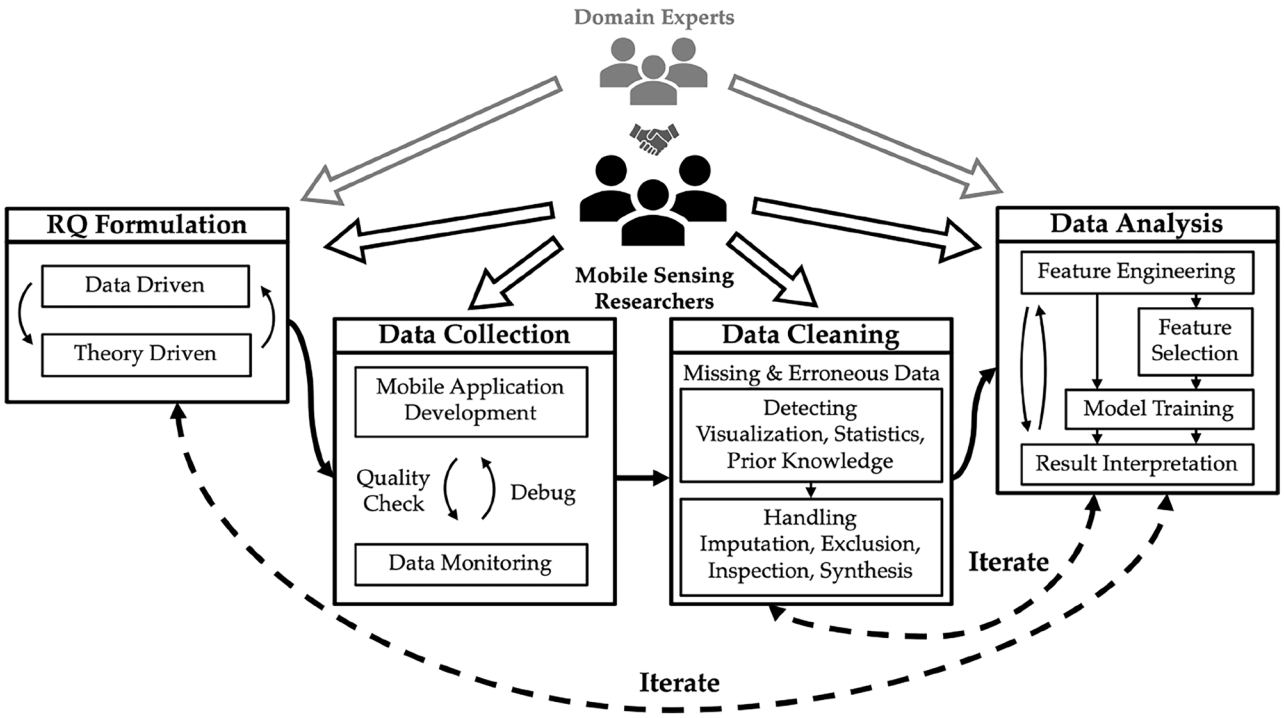


Fig. 1 The common practice pipeline when conducting passive mobile sensing research

3.4 Interview analysis methods

We used DeDoose[1] to analyze the transcribed interview text and the interview notes. One author first employed an inductive coding strategy to develop a codebook. They extracted bullet points from each participant's interview data, as a summary of each participant's answers. Note that we intentionally summarized the answers for the practices and the challenges separately, so that we could have a clear focus when analyzing each part. Then, the author compared and merged the responses across participants, which generated an ordered list ranked by the degree of commonality, i.e., how many participants followed similar practices or encountered similar challenges in their research. Finally, the author discussed with the other authors, and iterated the codebook until consensus was reached.

The following sections summarize the results and findings of the interview analysis, from the perspective of common practices (Sect. 4), as well as pain points and needs (Sect. 5). Moreover, we also summarize the results of the last two questions about researchers' roles and the uniqueness of mobile sensing data in Sect. 6.

[1] https://www.dedoose.com/.

4 Results: common practices

In this section, we summarize the common practices that participants adopted during their research. Researchers with little or no previous experience can best benefit from this section as it provides an overall picture of the state-of-the-art research procedures in this area.

We first highlight the overall common pipeline adopted by participants in Sect. 4.1. We then present the results in the same order as the interview protocol, and summarize detailed common practices in the pipeline, including data collection (Sect. 4.2), data cleaning (Sect. 4.3), and data analysis (Sect. 4.4).

4.1 General pipeline

All of our participants have a technical education background. However, their high-level research goals were usually about human behavior such as mental health (P1, P3) and digital device usage (P4). Thus, collaboration with experts in a specific domain area is a very common practice. All participants established different levels of collaboration with one or more domain experts, and worked jointly with these experts to answer their research questions.

Overall, participants had a similar procedure throughout their research projects. Four major steps were involved: research question formulation, data collection (if needed), data cleaning, and data analysis (see Fig. 1).

In the formulation step, participants mainly took two types of approaches: theory-driven and data-driven. Participants (six out of ten) often resorted to literature and their collaborators (i.e., domain-experts) before proposing their research questions. Literature and expertise provided well-established theories that could guide their research. Meanwhile, because of the richness of mobile sensing data, participants (seven out of ten) also used a data-driven approach. Before settling on a research question, they collected and investigated a dataset in an exploratory way to look for interesting questions to ask. "*Let the data speak!*" (P3, P8). Participants either focused on one method or leveraged both methods jointly to frame their research questions.

Although there are a few public passive mobile sensing repositories (e.g., (Wang et al. 2014)), their coverage is limited and many research questions require more specialized datasets. Therefore, researchers often collect their own datasets. Three participants used existing datasets collected by other research groups, while the remaining seven participants all conducted their own in the wild data collection studies. Nevertheless, regardless of whether a public dataset or a new dataset was used, data cleaning is a necessity after having the raw data, as mentioned by all participants. Then, data analysis is conducted on the cleaned data, during which domain-experts are often involved for feature engineering and results interpretation (as shown in the grey arrows in Fig. 1). It is noteworthy that sometimes research questions appeared to researchers not at the beginning of their work, but only after exploration of their dataset. For example, P9 published a paper about daily habit behavior mining, but the focus on daily habits was not determined at the beginning. "*Initially, we did not have a clear research question. We mined the rules after we got the data and found some interesting rules, then the research questions became clear.*" (P9) This shows that there can be an interesting iteration between data analysis and research question formulation (P3, P6, P7, P9), as shown in Fig. 1. In addition, we point out another common iteration between data cleaning and analysis. Participants reported that it was often hard to complete data cleaning all at once because of the richness of the mobile sensing data. Participants mentioned an iterative process between the two steps. "*We did some cleaning, and then we analyzed a bit... Then we cleaned our data in a better way.*" (P8) Researchers often realized how they could improve data cleaning after some initial analysis, so they went back and modified the cleaning step, and use the refined data for further analysis.

4.2 Data collection: monitoring

Seven out of ten participants ran in-the-wild studies to collect passive sensing data. In order to obtain data from sensors in mobile phones and wearable devices, developing a mobile application has become the standard way. Three of the seven participants developed mobile applications based on the AWARE Framework[2] (Ferreira et al. 2015). One participant used an internal application developed by a private company. The rest developed their own mobile applications for data collection.

Ensuring high data quality is one of the most important goals of data collection studies (Hernández et al. 2017; Ferreira and Ferreira 2017). Among the seven participants, six used monitoring methods during their studies. Data streams from multiple sensors on mobile phones and wearable devices (if any) are uploaded to a server and the size of the data is usually huge. Therefore, monitoring the incoming data is important for researchers to keep track of the study status. A dashboard that visualizes the status of incoming data streams (e.g., amount and quality) was commonly adopted by the majority of the participants (five out of six). For instance, P10 ran a study to collect GPS, WiFi, Bluetooth data from smartphones, and physical activity data from smart bands. "*We developed a dashboard to visualize the amount of data from each sensor of each participant every day. This helped us a lot! We knew the cases where some sensors' data did not come in.*" (P10) Other common methods include periodically querying study databases (P1, P2, P6), and checking whether new data files were generated (P5, P7). For example, P2 set up an automatic process on the server to monitor the data collection study. "*We set up cronjobs to do automatic querying once every few hours to get a sense of the current data collection progress.*" (P2) When the monitoring process detects abnormal data behavior (e.g., some sensors' data are missing), researchers can quickly react to it. Researchers often start with debugging the server or the mobile application, as reflected in the loop within the data collection block in Fig. 1. They also reach out to users when necessary. For example, P10 ran into a hardware issue and the server stopped collecting data from users' phones. After they fixed the problem, "*...some sensors' data did not come in. Then we followed up with our participants [to get a copy of their local data].*" (P10) The team found that the only way to collect the missing data was to ask users to send data that was being cached on the phone. Although such an effort was time-consuming, this helped to maintain the integrity of the data.

4.3 Data cleaning: problematic data

Due to frequent hardware and software issues (Matarazzo and Pakzad 2016; Wang et al. 2016a; Zhou et al. 2018), missing data and erroneous data are inevitable during data collection. Therefore, data cleaning is important to perform

[2] https://awareframework.com/.

before doing any meaningful analysis. Participants used various methods to detect (Sect. 4.3.1) and handle (Sect. 4.3.2) problematic data. We highlight the common practices in these two steps.

4.3.1 Detecting—visualization, statistics, and prior knowledge

Visualization and statistical tools are methods widely used in a typical data cleaning process (Bilogur 2018; Valero-Mora et al. 2019), and mobile sensing data studies are no exception. Six participants used these methods to identify missing or erroneous data. For visualization, histograms (P1, P3, P6) and time-series line plots (P1, P2) were frequently mentioned. Participants also use simple statistics, such as direct data counts for missing data (P1, P2), confidence intervals (P4, P6, P7), duplication tests (P2, P3), and normality tests (P6, P7), to detect missing value and abnormal data. These methods are not specific to mobile sensing data.

Prior knowledge of the data also played a significant role in problematic data detection. The passively sensed data capture users' daily behavior and researchers had certain expectations on how a user's data should look like, based on their own commonsense and life experiences. If the visualization or statistics of the data deviates from expectations, then it suggests the possibility of abnormal data. For instance, the overall daily step count should fall into certain ranges, such as from 0 to 50,000 and the step count is time-dependent. With such an expectation, P3 found some abnormal step data. "*There was some [data] like one person having zero steps during the night, but suddenly he had 100 steps at like 3:10 AM. A weird small peak. This was definitely an outlier and it should be filtered.*" (P3) Another example is the location. P10 found that some users' location data did not make sense. "*You know, GPS can be very inaccurate. I have seen data showing this person is in a lake and obviously, they were not on a boat or swimming...*" (P10) In these cases, participants relied on their prior knowledge to recognize problematic data that won't necessarily be detected by statistical analysis.

4.3.2 Handling—imputation, exclusion, inspection, and synthesis

In data science, imputation and exclusion are the most popular methods for missing data and erroneous data, respectively (Zhang 2016; Pedersen et al. 2017). These methods are also very commonly used in the mobile sensing area. Seven participants used a range of methods to impute missing data, including previously last-seen imputation (P1, P3, P6, P7), average imputation (P1, P4, P9), nearest-neighbor imputation (P3, P8), etc. As for the erroneous data, participants usually excluded them from the dataset. P10 particularly pointed out the potential value of inspecting the data closely because some meaningful outliers (e.g., intentionally powering off the phone might reflect the mental state of social distancing) could easily be mistakenly recognized as problematic data.

One interesting method somewhat unique to mobile sensing data is that some sensors can complement each other because sensors capture different aspects of information in the same context. Both P1 and P6 synthesized GPS and WiFi data to obtain better location information when one of them was missing (WiFi SSID could be used to estimate physical location given the WiFi geolocation information). P5 synthesized weather and users' comments to impute missing GPS data, e.g., if the weather was cold and a user claimed to be at home on particular days, locations were imputed with the home position. Multiple sensors capture the same environment from their own aspects. Therefore, these sensors can corroborate each other, allowing researchers to impute their data in a valid way.

4.4 Data analysis: from feature to interpretation

After cleaning the data, researchers conduct a variety of analysis steps to address their research questions. This is usually the core part of their research. Unsurprisingly, similar visualization techniques and statistics to those used in data cleaning (as summarized in Sect. 4.3.1), are used in data analysis. Beyond this, we further summarize other interesting analysis practices.

Other than activity recognition, most longitudinal in-the-wild studies have challenges in collecting valid ground truth labels, thus the number of data labels is often limited (e.g., (Wang et al. 2015; Xu et al. 2019)). Therefore, feature extraction is a common first step during data analysis after the data has been cleaned (Sect. 4.4.1). Sometimes, feature selection and model selection are needed, followed by the feature extraction (Sect. 4.4.2). In the end, researchers usually dive deep into their analysis to obtain meaningful interpretation (Sect. 4.4.3).

4.4.1 Feature engineering strategy

Given the data collected from a number of sensors, researchers need to decide what features to extract in order to obtain meaningful information from the raw data. Participants mentioned three different strategies to tackle this issue: referring to literature, consulting collaborators, and being creative.

Literature Referring to related work is one of the most common strategies to get inspiration for feature extraction. Participants referred to two types of literature to obtain guidance: existing mobile sensing literature and domain-specific literature. With the increase in use of passive mobile sensing in recent years, there is an emerging literature of human

state modeling with mobile sensing data across a range of domains. Participants were easily able to find existing mobile sensing publications in related areas. Seven participants mentioned that they used the examples of effective features in existing papers and adapted these features for their own research projects. For example, P6 studied the relationship between mobile sensing features and depression. "*Many of our location features were learned from the famous Saeb's paper*" (P1) Saeb et al. (2015). In addition to using existing research in the mobile sensing area, participants also searched for literature in the domain that was specific to their research goals. This was adopted by six participants. Continuing P6's example, "*Other than reading papers in our area, I also read a few psychology and psychiatry papers to see whether I could get any inspiration. And I did! I found that cyclical sleep patterns would reflect the influence of mental health. So I extracted this as a feature.*" (P1) These domain-specific papers provided theoretical foundations, which could guide specific features that align with existing theories.

Expertise from domain experts As shown in Fig. 1, data analysis is another major step where domain-experts get deeply involved. "*[We] talked with experts with research questions to tweak our analysis methods.*" (P10) Domain experts are familiar with the mediators and moderators that may influence behavioral outcomes of interest. However, they are often less familiar with technical details about the format and content of the data. On the contrary, mobile sensing researchers are more familiar with the data but have relatively less experience in the specific domain of interest. Interdisciplinary collaboration introduces expertise from both sides, which can point to directions for new feature extraction (P4, P8, P9). Specifically, participants often summarized the information embedded in the data so that it was straightforward and easy to understand. Then, they gave a high-level summary to their domain-specific collaborators. Experts commented on the existing features and provided ideas of new potential behavior features that could be captured by mobile/wearable devices. Since sometimes not all features were feasible (e.g., number of people that a user talked to, the amount of calorie consumed in a day), participants then created feature extraction code for the feasible ones. For instance, P7 worked with doctors to understand patients' behavior captured by mobile phones. "*When I was writing the code for feature extraction, I talked with my collaborator Dr.[anonymized] once a week to make sure these features make sense. I explained these features to him and got feedback. Sometimes he pointed out mistakes I made.*" (P7) Continuous, iterative communication between mobile sensing researchers and domain experts not only benefits effective feature extraction, but also eliminates potential misunderstanding and increases the likelihood of obtaining a valid result.

Creativity As mobile sensing data capture users' daily behaviors at a fine resolution, it is common for researchers to create some features that are not mentioned in the previous literature and that domain experts might not think of. For example, theories suggests a positive relationship between more exercise and better mental health (Deslandes et al. 2009; Morgan and Goldston 1987). These theories do not involve a particular time or context of exercising, but mobile sensing can capture these relationships in detail, e.g., a user often talks with someone else over the phone while exercising. A corresponding feature can be "the duration of conversation during exercise" but no existing theory would have directly suggested this feature. Five participants used their own life experiences and intuitions when designing new features for extraction. "*Some features were extracted by the essence of the research question, intuitively.*" (P1) Although there was no direct supportive literature, P1 extracted communication features specific to users' family and close friends (e.g., number/duration of phone calls). In another example, besides consulting experts, P9 further relied on their own life experiences to extract additional location features. "*We also used some rule-based heuristics to get features such as 'from work to home'.*" (P9) These creative features proposed by participants themselves were effective in their own projects.

4.4.2 Selecting the right features and models

After feature extraction, some participants further conducted feature selection for the purpose of their research goals. This was usually due to the large number of features compared to the data sample size or the colinearity among features (P1, P4). They used common techniques including visualization (P4, P9), information gain (P1, P6), gini index (P6), correlation analysis (P1, P7), principal component analysis (P7), and so on.

Six participants' projects involved typical machine learning classification or regression tasks as one of their research goals. Off-the-shelf traditional machine learning models were used, such as linear mixed model (P2), random forest (P1, P4), extreme gradient boost (P3, P8), and support vector machine (P1, P6). Participants particularly mentioned not using deep learning because of the lack of data and the difficulties in interpretation. This is also reflected in a large amount of existing literature (e.g., (Mirjafari et al. 2019; Obuchi et al. 2020; Wang et al. 2018; Srinivasan et al. 2018)). When determining the specific model to use, participants generally tried a variety of machine learning models and chose the one with the best performance, using tools like Weka (Hall et al. 2009) (P4, P6) and packages such as Scikit-learn (Pedregosa et al. 2011) (P1, P2, P4, P5, P8, P9).

4.4.3 Results interpretation

Both feature selection and model training are optional processes. Some researchers may skip feature selection, or even skip both, as indicated by the arrows in the data analysis block in Fig. 1. But the analysis always ends with results interpretation, either directly on extracted features, such as correlation analysis between mobile features and behavior outcomes (P1, P6, P9), or on the trained models, such as feature importance comparison or instance inspection (P2, P6, P10). Researchers often resort to literature to triangulate and understand their results. Beyond this, similar to the feature engineering step (Sect. 4.4.1), domain-experts also get deeply involved in the process of the interpretation. These two steps are closely linked. The extracted features used for analysis (directly or through models) are the basis of the interpretation. As a proxy of human behavior, mobile features may not reflect a person's behavior directly. Participants often summarized their low-level analysis results into high-level understandings and discussed with their collaborators. For example, P9 worked on a project about investigating daily routine behavior from a user group. He extracted group behavior rules from a cleaned mobile sensing dataset his team collected, and then leveraged domain experts to get a better understanding of the rules. "*After getting the measure of [the rules'] group-fit, I started to cross-validate with experts to interpret the results.*" (P9) Moreover, there is an interesting loop between feature engineering and result interpretation. Continuing the previous example, P9 realized a huge number of ordinary, uninformative rules after consulting with experts. "*... so we go back to the mining algorithm and adjust the threshold [to obtain different rules].*" (P9) Feature extraction establishes the basis of the interpretation, and the interpretation can in turn inform adjustment or additional directions for feature extraction.

Beyond the loop within the feature analysis, participants also mentioned iterations between the data analysis and the research question formulation, as well as between the analysis step and the cleaning step (as described in Sect. 4.1). Due to the large number of sensors and the richness of each sensor's data, it is difficult for researchers to make a one-time pass on the overall pipeline. More importantly, these iterations are usually necessary before meaningful findings come out.

5 Results: challenges and needs

Knowing the common practices adopted by researchers can serve as an overall introduction and guidance for researchers who have interest but no previous experience in this area. In addition, it is important to understand the challenges and needs that researchers have during their research. This can not only provide more comprehensive guidance for laymen and beginning researchers, but also offer deep insights for intermediate researchers who already have some experience and are actively improving their skills.

We first summarize the results of the common challenges that stood out during participants' research, following the order of data collection (Sect. 5.1), data cleaning (Sect. 5.2), and data analysis (Sect. 5.3).

5.1 Challenges of data collection

Seven participants conducted in-the-wild studies to collect data with their own research team. They ran into several challenges during the process, including software development (Sect. 5.1.1), problem debugging (Sect. 5.1.2), communication with users (Sect. 5.1.3), and researchers' own stress (Sect. 5.1.4).

5.1.1 Efforts in mobile application development

As mentioned in Sect. 4.2, participants resorted to various methods to develop mobile applications. However, from their experience, no matter which method participants used, this procedure was quite time-consuming. "*One pain-point is the app development, [it] took too much time.*" (P2) Even with the help of an existing framework, participants spent extra effort to customize the app to their specific studies. For example, P6 added a graphics layer on the top of the AWARE framework in order to prevent users from changing sensor settings in their local application (a feature enabled by the original framework).

Different platforms required different data collection applications. Android and iOS are the two most popular platforms. Developing applications for both platforms doubles the workload. Therefore, some participants had to focus on only one platform (P5, P10). In particular, participants complained about the difficulties of getting data from iOS devices. The iOS platform imposes strong constraints on what an application can collect so it was hard to collect data such as application history even though researchers had consent from users. When asked about why they did not collect data from iOS users, P5 replied with what appeared to us as a bitter smile. "*It's painful to learn to develop an app on a completely different platform. But a more important thing is that iOS won't give me the data I want. I need to get the foreground app that users were using. This is impossible on iOS... so it doesn't make sense to collect data from iOS users*" (P5). Moreover, updates of mobile operating systems could also create problems as related privacy policies changed. "*The updates of the operating system can sometimes ruin the app. Some APIs were deprecated so I need to rewrite a few modules.*" (P1) For example, Android 9.0 started to restrict access to the microphone, camera, and

other sensors when an application is in idle mode or running in the background. A persistent notification will now appear on the phone if the application needs to access those sensors (and 2019). A change like this could significantly affect user experience during data collection studies.

Researchers also had a hard time improving battery consumption (Pérez-Torres et al. 2016; Min et al. 2016). Turning on high-frequency sensors, such as the gyroscope and microphone, significantly increases battery drain. P10's team spent effort optimizing the location data collection. "*The location sensor is energy-consuming. We spent a lot of time to optimize the battery life to minimize the effect on users' phones.*" (P10) A high energy consumption not only influences the data collection process (e.g., the phone dies more quickly), but also creates a negative user experience, which may affect participation compliance. Overall, researchers shared similar thoughts about application development difficulties. Building upon an existing framework could reduce the effort, but it was still demanding to build data collection applications.

5.1.2 Problem identification and debugging

As summarized in the common practices of data cleaning (Sect. 4.3), the hardware issue with sensors and servers, and software bugs in mobile applications are inevitable after a study is launched, no matter how well the study is designed and planned. This leads to problems of missing data, erroneous data, or even potential damage to the phone. Many participants actively monitored their studies, as indicated in Sect. 4.2. Therefore, they could quickly notice that there were some problems that were affecting the data collection. Identifying the causes and figuring out how to fix the problems while a study was running were important for ensuring the quality of the data. However, the process was often difficult and both time- and energy-consuming. Sometimes, even just identifying the origin of a problem took a long time (P1, P7, P10). "*The biggest struggle was finding new bugs and debugging. Throwing the data away is easy. It is more important but also more challenging to solve the problem from the root.*" (P4) Common issues include server database crashes, connection issues between servers and phones, application storage bugs, application sampling bugs, etc. However, participants often encountered unexpected bugs from the phones or servers and expended huge amounts of effort to fix them (P1, P2, P4, P6). "*[I] tried to anticipate as much as I could, but always, there were some bugs not foreseen.*" (P7) This was a big challenge for participants.

5.1.3 Difficulties of user communication

Sometimes, a problem cannot be completely solved by researchers themselves but requires coordinating with the users they are collecting data from. There are many problems that are related to users' behaviors. For example, users may turn off certain sensors to extend battery life and forget to turn them on later. These sensors' data will then be missing. In these situations, following up with users is often unavoidable. Communication with users can help researchers to (1) obtain a better knowledge of users' behavior such as forgetting to charge wearable devices (P2, P7), or (2) fix hardware/software issues (P1, P6), e.g., granting permission to or turning on certain sensors (P2, P10), or (3) collecting local data if the mobile application loses connection to the server (P2, P6), or (4) collecting additional information such as follow up questionnaires (P6). However, participants found such communication with users difficult and demanding. P7's study used a rolling-enrollment approach, where participants joined/left the study at different times. "*Participants were at different paces. Some joined earlier. Some joined later. Communication with them is difficult and hard to manage.*" (P7) The communication becomes more challenging as the number of participants increases. "*I prepared some email templates but I still had to customize some.*" (P4) P6 found that her users were particularly concerned about privacy and reluctant to answer her follow-up surveys (they signed the consent but changed their minds). "*Privacy is a big concern when we tried to reach out to them for more survey questions.*" (P6) There were even cases when users completely disconnected from the research team. "*They just didn't reply... Sometimes they got back to me a few weeks later. What can I say?*" (P5) All these problems forced participants to spend extra time on managing user communication. Some participants' teams recruited study coordinators to help (P7, P10), but participants still needed to identify those users' problems to decide what to communicate. The ultimate goal of passive mobile sensing is to completely remove user burden and *seamlessly* provide personalized services to facilitate users' life experience. Nevertheless, the current state-of-the-art of passive sensing is still in the research stage and is not completely user-burden-free yet. Communication with participants is inevitable and it requires energy and time from researchers.

5.1.4 High stress of researchers

One interesting thing to highlight was that a few participants described themselves as being highly stressed during the study (P1, P2, P4, P5, P6). For example, P5 worked on a project involving intelligent interventions based on detected user behavior. "*I was very stressed when the study was launched at that time. The study cost my supervisor thousands of dollars and I really wanted to get the data in a good shape. Otherwise, it would be a waste of time and money.*" (P5) Participants reported a variety of stressors including the data quality, the high cost of the study, unexpected technical

issues during the data collection, etc. These stressors kept participants tense during the data collection period. When reflecting on their data collection studies, participants had mixed feelings. Some participants thought that being stressed was fine (P2, P7). "*This was a big motivation for me.*" (P4). While some thought they were stressed for no reason. "*There was no need for being so stressed out... my worries won't change anything.*" (P5) However, no matter what thoughts they had, participants agreed that staying alert and responding in a timely manner to issues were necessary for obtaining high quality data.

5.2 Challenges of data cleaning

In spite of efforts in data monitoring (Sect. 4.2), application development (Sect. 5.1.1), and intensive debugging (Sect. 5.1.2), the issue of problematic data is inevitable. This leads to challenges in the data cleaning process (Sect. 5.2.1). Moreover, due to the heterogeneity of the data, how to leverage and reuse existing code also becomes a problem (Sect. 5.2.2).

5.2.1 The lack of objectivity, interpretability, and verifiability

As illustrated in Sect. 4.3, using standard visualization and statistical methods to detect and handle problematic data are common practices of data cleaning. Moreover, as the mobile sensing data are closely connected to daily routines, researchers can leverage intuitions on what the data should look like, which makes the data cleaning easier. However, participants were not satisfied with their data cleaning. There were three main perspectives: objectivity, interpretability, and verifiability.

(1) The lack of objectivity. Although using standard data science methods, participants still blamed themselves for making subjective decisions when cleaning the mobile data, such as setting a decision threshold (P3, P5, P8). "*We had certain decisions with respect to the threshold of missing data [for excluding a feature or a participant], but the decision was too subjective.*" (P5) (2) The lack of interpretability. Participants found it hard to understand the actual reasons for missing or erroneous data (P7, P10). Being able to interpret the reason would help guide the cleaning. "*To detect potential outliers, [I] have to know what causes it and what it looks like... There was no ground truth for finding outliers. It was only based on estimated guesses.*" (P7) When cleaning his data, P7 used established methods such as Z-score for outlier detection and DBSCAN for imputation. But he did not feel secure, as these methods did not tell him what caused the outliers or the missing data. And (3) the lack of verifiability. This was related to objectivity and interpretability. Without knowing the actual user behavior, it was difficult to evaluate the validity of data cleaning (P4, P5, P7). For example, P7 found it hard to determine whether an excluded data sample was actually erroneous or not. "*The algorithm [Z-score] found a period with abnormal high physical activity, but maybe this user was doing some extreme exercise? I don't know! The data did deviate from the majority but not too much.*" (P7) Due to these reasons, the cleaning process was still regarded as challenging in spite of the close connection between mobile data and real-life experience.

5.2.2 Extra work for multiple sensors and datasets

Different sensors have different properties and formats, which need to be carefully taken care of when cleaning data. For example, P6's dataset was collected from college students. When she tried to clean the phone call data, she found that the data was sparse and many days' data was missing. "*Then I realized that even I often have zero calls in a day. Young people get connected through social platforms...*" (P6) Therefore, she did not do any imputation on the call data, while she did use standard imputation methods on other sensors' data such as step counts. Participants could not simply use the same piece of code to deal with multiple sensors. Instead, they developed cleaning code for each sensor separately (P1, P10). This requires extra work.

Moreover, although the interview mostly focused on one project chosen by participants, they also mentioned the challenges beyond the single project when they were asked about their pain points. Eight out of ten participants were involved in multiple projects with different data collection studies. However, different studies had different settings, with different sensors of different sampling rates (Blunck et al. 2013). These differences made it hard for participants to have cleaning scripts that could handle multiple settings (P1, P3, P6, P9, P10). "*There were too much data from different sensors that sometimes I was not aware of the error in my code. This became worse when I was cleaning multiple datasets because the data were different. The code worked on the first wouldn't work on the other. So I ended up writing multiple repositories for the cleaning.*" (P9) Participants paid extra effort to develop ad-hoc cleaning scripts and check the quality of the cleaned data. Beyond data cleaning, data heterogeneity further brings up more problems in the analysis step, as shown in Sect. 5.3.3.

Confronted with these challenges, five participants mentioned the need for a tool to ease the process of data cleaning, to address issues such as high-missing-rate warning (P4), automatic imputation (P1, P3), and outlier detection (P2, P6, P8), etc. P1 first expressed the need for an automatic process to address the missing data problem. "*It would save my days if the software can provide the optimal way of handling missing data.*" (P1) After considering different approaches, she took a step back from a completely automatic process and

thought about leaving space for manual decisions. "*I guess it cannot have too much imputing, otherwise it will bring in too much fake data that biases the final results of the model. This can be a research topic itself. [smile] Maybe a tool that can provide multiple options to process missing data is good enough.*" (P1) Support for manual decision making was echoed by other participants. They did not expect the tool to completely address the problematic data in an objective and reliable way. Instead, they wanted the tool to provide data cleaning functions, but at the same time, the tool should allow them to have control of the cleaning process. "*An outlier detection tool is important. But there is no one-hundred-percent 'correct' way to detect outliers. Maybe the tool can detect some high-probability outliers and show them so I can take a look.*" (P8) Participants would like the tool to provide flexible assistance with cleaning so that they could make final decisions. We will have more discussion on researchers' desire for full control in Sect. 6.1.1.

5.3 Challenges of data analysis

In the data cleaning step, we showed that dealing with problematic data is not only a common practice (Sect. 4.3.2) but also a challenge (Sect. 5.2.1). We find a similar case in the data analysis procedure as well: both model selection and results interpretation are not only the common practices (Sect. 4.4.2, 4.4.3), but also the two main challenges (Sect. 5.3.1, 5.3.2). In addition, issues of data heterogeneity (Sect. 5.3.3) and communication with experts (Sect. 5.3.4) are also frequently encountered by researchers.

5.3.1 Great efforts to pick the right features, models and parameters

Because of the large number of extracted features commonly found in passive sensing studies, participants spent a large amount of time on testing and comparison, in order to find the most effective features, best algorithms, and the appropriate parameters that result in the best performance. This was regarded as one of the major challenges in data analysis. Five participants mentioned that trials of various algorithms for feature selection and model selection were time-consuming. For example, P6 needed to select from over six hundred mobile features. She tried both information gain and gini index as the selection criteria, which led to two close but different feature sets. Both methods were valid and supported by statistics and machine learning literature (Manek et al. 2017; Lee and Lee 2006). "*I couldn't tell which one worked better by just eyeballing. Both feature sets made sense.*" (P6) She ended up comparing the performance of the two models trained on the two sets by going through the whole training and tuning pipeline, which was time-consuming. P8 had another example of model comparison, where he had to spend extra effort on writing model code from scratch. "*If the models you'd use are all supported by scikit-learn [a Python package] then things are easy. But things were different in my project. I had to re-implement the models cuz I couldn't find any existing implementation.*" (P8) It became more demanding when parameter tuning was taken into account. Many participants' machine learning models involved parameters that required extra tuning time. Three participants mentioned that parameter tuning was tedious. "*We struggled at manually tuning the hyperparameters at the beginning... We developed our own automatic parameter tuning pipeline. But the training still took long.*" (P4) Note that the challenges of feature selection, model selection, and parameter tuning stem from the areas of data science and machine learning rather than mobile sensing itself. Improving skills in these areas or collaborating with experts in machine learning may help researchers more easily tackle these problems. Six researchers also mentioned the need for a tool to help with comparing and tuning models,

An automated tool for features, models, and parameters searching could greatly facilitate the research process. "*After setting up the input features, the tool can select the algorithms and tune the parameter automatically.*" (P6) Similar to Sect. 5.2.2, the majority of the participants wanted the selection and tuning procedures to be available and transparent so that they could control them and make decisions on the final choices of features, models, and parameters. "*A tool that can automatically select features and machine learning models and tune the hyper-parameters. It's like a full system but gives the option to make some changes within a certain part of the pipeline and each component should be flexible.*" (P5) Such a tool would allow researchers to focus on understanding the data and interpreting the results, which is an essential part of mobile sensing research (Sect. 6.1.2).

5.3.2 Vague evaluation and interpretation

Some research questions in passive mobile sensing have recognition and prediction tasks that can be directly evaluated, such as depression detection (Xu et al. 2019; Lu et al. 2018), emotion recognition (Zhang et al. 2018; Mehrotra et al. 2017), and application usage prediction (Wang et al. 2019b; Chen et al. 2019). However, there are also research questions that lack standard evaluation metrics, such as the quality of behavior rules. Participants mentioned that they had difficulties in defining their performance metrics (P5, P8, P9). Take P9's routine behavior mining as an example. "*We mined a set of behavior rules and they made sense to us. However, we had no idea how to evaluate their validity. We could only resort to users but that might easily be biased.*" (P9) Finally, the team conducted interviews with users to evaluate the effectiveness of the rules. When lacking an objective metric like classification accuracy or regression

error, participants worried that the evaluations were not conclusive enough.

Moreover, mobile sensing data are only a proxy for human behavior. The information captured in mobile sensing data is ambiguous. Therefore, results interpretation is often not straightforward. Participants explicitly mentioned their hard time developing appropriate interpretations of the results (P4, P9, P10). For example, P4 developed an unsupervised clustering algorithm to understand mobile phone usage behavior patterns. The algorithm output a large number of clusters, but it did not help to answer how one usage pattern cluster was different from others. Interpreting the clusters required P4 to look deep into each cluster, summarize the difference between clusters, and generate human-understandable findings. "*Using our algorithm, we obtained clear clusters of users' app usage behavior. But I spent over one month to understand and make sense of these clusters.*" (P4) She further tried to triangulate her findings by conducting user interviews. But the interview sample was a very small fraction of the mobile data sample. "*Some participants gave me feedback that was contradictory to what they did on their phone... The data only reflected 'when' and 'which' mobile application users used, but did not tell us 'why'.*" (P4) As mentioned in Sect. 4.4.3, the use of literature and experts often help with the interpretation. However, due to the fine-grained data resolution provided by mobile sensing, sometimes findings are beyond the coverage of existing literature. This is also reflected in the feature engineering step (Sect. 4.4.1). "*I found some really interesting behavior patterns in the data. But I did not find any papers to support my findings. I consulted a few experts in behavior science but did not get a satisfying answer.*" (P9) In the final version of the publication, P9 presented these findings, provided shallow discussion, and encouraged domain-experts to explore this in future work. Both examples illustrate the difficulties that mobile sensing researchers face in trying to understand their results.

5.3.3 Data heterogeneity and low code re-usability

Similar to the challenge mentioned in data cleaning (Sect. 5.2.2), the heterogeneity of the data also causes difficulties in data analysis. Within the same dataset, developing an analysis pipeline that can be re-used for data from different sensors, formats, and platforms is difficult. Some mobile sensors are event-based (e.g., phone call) while others are sample-based (e.g., GPS location). Calculating the number of phone calls and the number of places that a user visits are completely different (P1, P2, P7). Even within the same feature type, only generic aggregation functions (e.g., calculating the minimum and the maximum) are common. In order to generate more specific features, participants had to develop individual feature engineering code for each sensor (P3, P6, P7).

Moreover, when multiple platforms are involved in one study, the differences between platforms are reflected in the data format. "*There is a large variability in the data format. I iterated my code many times to deal with different mobile platforms.*" (P7) For instance, the Bluetooth sensor from most Android devices usually provides the MAC address of the scanned Bluetooth sensors, but this is usually not available for iOS devices. Dealing with these differences hindered participants in re-using their code and required more programming effort.

Multiple datasets add another layer of heterogeneity. For example, P8 had a pipeline for an analysis algorithm. But the algorithm did not include any feature engineering and assumed well-structured feature inputs. So he had to unify the data from different studies before applying the algorithm. "*Although my team had the code of the algorithm, it takes a standardized input at a feature level. I have multiple datasets with different data policies and raw formats... Preparing them for the algorithm is laborious.*" (P8) Overall, due to data heterogeneity, participants found it hard to re-use their code for multiple purposes. They had to establish specific pipelines for different parts of the dataset (P3, P4, P7, P10) and different datasets (P8, P9).

Given the rich literature on mobile sensing in recent years, a large number of mobile features have been proven to be effective for various research questions, such as location-related features (Gonzalez et al. 2008; Imai et al. 2018; Canzian and Musolesi 2015; Xu et al. 2019) and physical activity features (Lane et al. 2011; Althoff et al. 2017; Doryab et al. 2019b; Aledavood et al. 2019; Xu et al. 2019). Five participants wanted a tool or a code library that could easily extract a set of predefined features from data so that they could quickly leverage these features. "*The location features, you know, log variance, entropy... They are useful. When writing the feature extraction code, I was wondering, there should have been some standard library for these common features.*" (P7) They also hoped that the tool would have the ability to automatically handle heterogeneous data such as different sensor settings and platform types. "*For example, leave them [the settings] as the parameters of a feature extraction function so it can easily process the heterogeneous data.*" (P4) Moreover, for specific research questions, the tool should support the use of new features that can be proposed based on the literature, suggestions from experts, and researchers' own creativity (Sect. 4.4.1). P5 mentioned the need for flexibility in a tool to add new customized features when needed. Being able to add new features is also important for researchers.

5.3.4 Difficulties of the communication with domain-specific experts

Interestingly, communication appears to be a challenge not only with users in the studies (Sect. 5.1.3), but also with domain expert collaborators on the research team. Four participants mentioned that communication with the domain-specific collaborators was time-consuming. One of the main factors is the delayed synchronization. If the loop between the feature engineering and result interpretation (i.e., the data analysis block in Fig. 1) involves both domain experts and mobile sensing researchers, a low-latency communication channel needs to be maintained among the team in order to keep everyone on the same page. For instance, P5's project was in collaboration with psychology researchers when designing behavior intervention techniques. "*It occurred that after the code was implemented, the psych team proposed some ideas so that we had to modify the code. It would have been better to have the conversation beforehand.*" (P5) Delayed communication introduced unnecessary iterations and cost additional time.

Another contributing factor is misunderstanding that can happen on both sides. Mobile sensing researchers can make mistakes when translating experts' advice into specific implementations. And, domain experts can misunderstand the actual meanings of mobile features and machine learning model outcomes. From our interview results, the latter appeared to be more common (P3, P7, P9). For example, P7 worked jointly with medical collaborators, who had limited technology backgrounds. "*Sometimes we had to spend extra time explaining the calculation to the doctors because of their misunderstanding, since they were not familiar with the technical details.*" (P7) Mobile sensing researchers need to ensure the synchronization and obviate misunderstanding on the team. An open, transparent communication loop is a necessity for such an interdisciplinary collaboration. With the level of detail and the richness of mobile sensing data, this becomes particularly important.

6 Results: the role of researchers and mobile sensing data

Sections 4 and 5 present participants' common practices, major challenges and needs during their research projects. In addition to knowing "what do researchers do" and "what do researchers need", we are further interested in understanding for which aspects they regard themselves as being essential to in the whole research procedure (Sect. 6.1). As seen in the common practices for data analysis (Sect. 4.4), a large number of the research questions can be framed as machine learning problems (such as regression or classification). Knowing the similarity and the difference between these mobile-sensing-specific problems and traditional machine learning problems can provide insights useful to passive mobile sensing (Sect. 6.2).

We envision the findings in this section inspiring a wide range of researchers with or without previous research experience, and provoke more reflections on the essential role of mobile sensing researchers and the uniqueness of mobile sensing data.

6.1 Essential role of researchers

After answering the help that participants wanted to get from a tool (Q.21), they were further asked about a question in the opposite direction (Q.22): what were the essential parts that should be conducted by researchers *rather than* an intelligent tool. Participants were asked to reflect on their projects and identify the parts that relied on their input. Two main themes emerged from the answers: manual control of the whole procedure (Sect. 6.1.1) and the interpretation of the results (Sect. 6.1.2).

6.1.1 Control of the procedure

Interestingly, although visualization, problematic data handling, feature extraction, models selection, and parameter tuning stood out as pain-points, and participants *did* want a tool to help them (Sects. 5.2.2, 5.3.1, 5.3.3), they emphasized the importance of having full control of the research procedure, from the data collection, data cleaning, to data analysis. They *did not* want an end-to-end black box. "*I do not trust a black box. Everything should be transparent in what it is doing!*" (P2) Participants stated that a transparent procedure and having tight control of it could leave space for human intelligence. "*Having the tool X to automatic extract feature would be good... but as for the modeling, [we] would like to control the step. Modeling is the interesting part.*" (P8) Even if having an automated tool helps to address the challenges described, participants wanted to have the power to inspect and change the procedure so that they could make important decisions (P4, P9):

> I did mention that I want the tool to help a lot. But it could not just do everything; otherwise, we researchers would become useless. [laughter] Maybe in customer products, the process could be automated, but that's not gonna happen in the near future. We are still in the early stage. The tool can help to process the data, but it is we who leverage our intelligence to make the decision. (P4)

As reflected from P4's quote, the decision making process was regarded as one of the main contributions of a publication. Researchers do not want to hand over this process to some tool, but would prefer a collaborative approach, i.e.,

the tool provides assistance for researchers to avoid mistakes and make better decisions. This coincides with the recent trend of human-AI collaboration (e.g., (Wang et al. 2019a)).

6.1.2 Data understanding and interpretation

Despite the fact that the difficulties of data interpretation were regarded as a big pain point and challenge (Sect. 5.3.2), participants had consensus that understanding the data and interpreting the results is one of the most important parts involving human intelligence. Especially for mobile data where the relationships between sensors are interleaved and complex, human knowledge from mobile sensing researchers and domain experts is the key to moving from data and models to final outcomes. This was also acknowledged by participants as the main contribution of researchers. "*The focus of the researchers is to find some interesting conclusions from the data. This is the most interesting part, right? Readers often don't care how you clean the data or extract features. They care about what you find eventually. So you need to really understand your data to dig out interesting findings.*" (P6) Participants trust themselves more than a tool for understanding and interpretation (P1, P6, P7). "*I don't think a machine can be intelligent enough for interpretation.*" (P7) They were also worried about the bias introduced by the tool (P2, P9, P10). "*Researchers need to have a good familiarity with the dataset, otherwise [they] might be misled by visualization or algorithms.*" (P2) If there existed such a powerful tool, most participants would leverage it up until the interpretation step of the results and leave this essential step for themselves.

6.2 Uniqueness of mobile sensing data

At the end of the interview, participants were asked how the mobile sensing data involved in their projects differed from other machine learning datasets. Mobile sensing data are a collection of longitudinal data from multiple mobile sensors that capture human behavior. Therefore, participants categorized it as a type of multi-dimensional time-series data (P1, P3, P5, P6). However, compared to other time-series data, mobile sensing data has its unique characteristics. Participants mentioned two aspects that complement each other: the data are information-rich but also information-vague.

On the one hand, passive mobile sensing data capture a wide range of users' unspoken behavior in the wild since users carry mobile devices and wearables almost every day to almost everywhere they go (Lane et al. 2010) (P1, P6, P10). With multiple sensor streams, the data are a continuous, longitudinal representation of human behavior from various perspectives. "*Although they are always messy, mobile sensors' data are very rich and contains much useful information.*" (P7) This makes mobile sensing data different from other time-series data such as inertial measurement unit (IMU) signals (Bulling et al. 2014) and electromyography (EMG) (Fan et al. 2018) which usually contain simple information of a single dimension. "*A lot of mental health symptoms are manifested by daily behavior, often in the long run. Mobile sensing can capture that. I don't think there exist other data that has a similar capability.*" (P8) The richness of people's daily behavior information in the data is the basis of any meaningful analysis and equips the mobile sensing technique with the capability of being applied to so many fields.

On the other hand, the information in the mobile sensing data is also vague. As a proxy for human daily behavior, it is hard for mobile data to capture and represent human behavior exactly (P4, P5, P7, P10). For instance, a combination of the GPS, WiFi localization, and ambient noise sensor could infer whether a user gets involved in an indoor activity (Wang et al. 2014), but it could not reflect the details of the activity or the exact behavior of the user. "*Mobile sensing is ambiguous. The information is latent and just a representation. It requires a lot of interpretation. You can only estimate but cannot have the exact information.*" (P8) Moreover, the relationship between the volume of data available and the amount of information can be mismatched. "*... mobile sensors' data are very rich... But the data size is not proportional to the amount of meaningful information. More data does not necessarily ensure more information.*" (P7) Because of the intrinsic complexity of human behavior, for classification and regression problems with mobile sensing data, sometimes even the labels may be inaccurate. For example, P3 used self-reported emotion scores collected by ecological momentary assessment (EMA). He believed that self-report is one of the most accurate ways to measure users' emotions because a person is the one who knows themselves best. However, he also had concerns because sometimes the results reported by users could be influenced by subtle factors such as environments in which users were completing EMAs, which could greatly affect the accuracy of their subjective feelings. This is different from other types of machine learning data such as computer vision or natural language processing where labels are usually accurate.

This interesting combination of being both information-rich but also information-vague highlights the uniqueness of mobile sensing data.

7 Discussion

Based on the results of the interviews, we discuss the implications of practices for researchers, especially for those who have interest but have little or no experience in passive mobile sensing (Sect. 7.1). We also discuss the design guidelines for a toolkit to help researchers during their studies, as

reflected from participants' challenges and needs (Sect. 7.2). Finally, we reflect on the limitations of our work and opportunities for future work (Sect. 7.4).

7.1 Practices for future researchers

The findings summarized in Sects. 4 and 5 can serve as good basis for practical guidance. We provide suggestions from four perspectives: combining the methodologies of data-driven and theory-driven (Sect. 7.1.1), monitoring data collection studies (Sect. 7.1.2), paying efforts on data cleaning and analysis (Sect. 7.1.4), and collaborating with domain-experts (Sect. 7.1.4).

7.1.1 Combining data-driven and theory-driven

In the projects chosen by participants for the interview, the majority of them (seven out of ten) mainly followed the data-driven methodology and the others started with the theory-driven method. However, these two methods are not exclusive. For instance, P4 started by extracting a number of behavior features from the data, which is a typical data-driven approach. She triangulated her results using the literature and additional interviews after she discovered something interesting. This follows a bottom-up path, where evidence from the data is used to support theories and literature. In contrast, the theory-driven method is a top-down path, where theories are leveraged to guide data processing and analysis. These two paths are compatible with each other. Researchers can leverage theories from the literature or experts to have a general direction for the research. Meanwhile, they should also explore the data to obtain intermediate results. The two approaches can be combined and used to support each other. Such a combination can leverage the value of theories to boost the efficiency of the analysis, and utilize the property of the data-driven method to have good coverage of the whole dataset, including aspects not covered by theories.

7.1.2 Monitoring to ensure data quality

If researchers plan to run their own studies to collect data in the wild, using a dashboard for monitoring is strongly encouraged. A visualization dashboard is a good tool that can facilitate the data collection process. It is more effective and intuitive than other methods such as direct database querying. Five out of seven participants employed a certain type of dashboard to monitor their own data collection studies. This greatly eases the burden for researchers to track the status of studies. Researchers should pay special attention to problematic data during the study, such as missing data and erroneous data, and respond instantly, e.g., fixing hardware and software bugs and contacting users if necessary. This can significantly improve the quality of the raw data, and ease data cleaning and the data analysis stages. However, researchers should also find a balance between the workload and their own mental health.

7.1.3 Leave time for data cleaning and analysis

Our interviews illustrated the importance of data cleaning and data analysis and the fact that they are demanding processes. Researchers should be patient during these stages, and pay special attention to the following aspects: problematic data handling, feature extraction, model selection, and parameter tuning. According to the participants in our interview, these parts are very time-consuming, sometimes even laborious and tedious. However, these steps are prerequisites before obtaining any valid scientific findings. Methods such as double-checking analysis scripts and creating unit tests (P2) are strongly encouraged during these processes. After getting all these prerequisites ready, data understanding and result interpretation are essential steps that researchers want to dive deep into. Researchers should establish a proper expectation of the effort required for data cleaning and data analysis, and leave enough time for these stages.

7.1.4 Close interdisciplinary collaboration with domain-experts

All participants in our interview are technology-focused. These researchers play important roles in system deployment, application development, data collection, cleaning, and analysis. Our interviews highlighted the role of domain experts in helping technically-focused researchers. Knowledge from domain experts can make a big difference at various stages during the research process: at the beginning of the study to point out directions to follow, in the middle of the iteration between theory and data to adjust the analysis, and also at the end when decoding the final outcomes to obtain meaningful interpretations. Having a close collaborative relationship with domain experts can help technically-focused researchers to obtain deeper and more valuable results. As summarized in Sect. 5.3.4, communication with experts could be difficult if not carefully attended to. Therefore, technically-focused researchers should have timely collaboration meetings with domain experts and provide detailed explanations of their work to avoid misunderstandings.

7.2 Design guidelines for a toolkit

Researchers wish to have a toolkit that can help them with data cleaning and analysis. Knowing the pain points and needs from researchers sheds light on the design of such a toolkit. It can provide common standardized computation

shortcuts, extract concise but meaningful information for stakeholders to quickly obtain insights from data, and help researchers to save time and energy so it can be applied on more creative analysis. We discuss three high-level design guidelines for such a toolkit, including sensor-modularization (Sect. 7.2.1), preliminary model tuning (Sect. 7.2.2), as well as flexibility and transparency (Sect. 7.2.3).

7.2.1 Sensor-based modularization

The issue of data heterogeneity is one of the biggest pain points, as indicated in Sects. 5.2.1 and 5.3.3. The toolkit should treat each sensor as a module that is independent of other sensors. Within each sensor, the toolkit should provide a list of predesigned functions for data cleaning (Sect. 5.2.2), visualization (Sects. 4.3, 4.4), and feature extraction (Sect. 5.3.3). These are mentioned by researchers as common practices or needs. Having these functions would greatly help researchers, allowing them to easily choose what support they want to use. Note that these functions can also be modularized, so that a function can easily be applied on multiple sensors when appropriate, to support optimum flexibility. For instance, using the toolkit, researchers would be able to easily modify the analysis pipeline by adjusting or changing modules for different studies, where they usually have different settings and collect data from different sensors. Meanwhile, the modularized design also supports expandability. The toolkit should have good coverage of sensors using a predefined sensor set. However, as new sensors emerge, the predefined set can be expanded. The toolkit would allow researchers to add new customized sensors as new modules when needed. Moreover, researchers can also easily add more ad-hoc functions (e.g., extracting new features) to certain sensors without affecting any other sensors.

7.2.2 Tuning of models and parameters

Participants have strong needs for support in feature selection, model selection, and parameter tuning (Sect. 5.3.1). Using a sensor-based modularized design, the toolkit should further provide a list of popular standard models, together with the function of feature selection and parameter tuning, so that researchers can easily conduct some preliminary tests. Packages such as scikit-learn could serve as a good starting point. The input of the model needs to be flexible, either using a single sensor or a customized sensor list. Each sensor's input feature list should also be easily modified to support selective training, where researchers can easily define the input as needed. Note that there is a trade-off between results and time. The more models and parameters the toolkit tests, the more complete the preliminary results will be, but the longer the selection and tuning will take. Therefore, the toolkit should allow participants to easily focus on a subset of feature lists, model lists, and the range of parameter tuning according to their needs. In addition, the toolkit should leave an open interface for researchers to add their own models. This can save them time when introducing their new models into the original pipeline. Such a design supports expandability and simplifies researchers' future work if the same models will be used again.

7.2.3 Retain the flexibility and the transparency

Participants emphasized the necessity to control the whole of cleaning and analysis (Sect. 6.1.1), and the toolkit should reflect this theme. Both the modularized framework and the automatic tuning design leave flexible spaces for researchers to have their own specific settings, e.g., picking certain sensors or models in one analysis and a different set for another study or analysis. These designs also allow researchers to add their own customized components into the research pipeline, e.g., including a new sensor or algorithm for analysis. Researchers should be able to easily control and adjust the process whenever they need to. Moreover, every component in the toolkit, such as the visualization functions, the feature extraction methods, and the specific machine learning models, should never be designed as a black-box. They should be completely transparent and provide descriptions as clear as possible, so that researchers can easily understand the detailed work being performed by the tool. This can avoid potential misunderstandings and mistakes, and leave it for researchers to determine whether they want to directly use the provided functions or develop their own customized components. By providing enough flexibility and transparency, the toolkit can give researchers complete control of the research process.

7.3 Different practices among researchers

In addition to the commonness among researchers, we also emphasize that the differences should not be neglected. The results from Sects. 4 to 6 mainly reflect the general practices, challenges, and needs. When conducting specific projects, researchers might follow a subset of the complete pipeline in Fig. 1. For example, P1 and P2 only used the theory-driven method when formulating their research questions (the leftmost part in Fig. 1), and P8 resorted to data directly. The rest of them leveraged a hybrid approach. Participants also employed different procedures in data cleaning and analysis, e.g., some selected features, while some did not. P6 also mentioned multiple projects she was involved in, where the specific practices also differed slightly. Overall, the specific practices and challenges vary and largely depend on the research questions that researchers aim to address. And our

results serve as a general reference for future researchers and software designers.

7.4 Limitations and future work

Our study has some limitations. First, the participants' backgrounds and coverage of the research application domains are limited. All 10 participants have a technical background, thus a large portion of the interview was focused on technical details and issues. None of the researchers come from a less-developed country. Moreover, there are some domains that none of the participants worked in, such as affective computing and activity recognition, where techniques such as deep learning are commonly involved. We plan to conduct more interviews with researchers from a wider range of backgrounds and domains in the future. Second, our interview approach mainly relies on participants' memory, assisted by their publications, if any. Some projects ended a few months before our interviews. This could potentially introduce bias during interviews. Instead, in the future, we could observe researchers during their analysis, or invite researchers to work on a standardized dataset. Third, there are many details that the interview did not cover. For instance, a few participants mentioned that they used different sensor settings in different studies. The interview did not go deep into inquiring about the decision process and their considerations. We plan to conduct more expansive interviews in our future work. Nonetheless, despite these limitations, our interviews paint a detailed picture of the process that passive mobile sensing researchers undertake when trying to understand human behavior.

8 Conclusion

In this paper, we address an important gap between the rich literature on contributions being made to particular application domains by using mobile sensing and the lack of focus on the researchers themselves and the processes they use. We conducted semi-structured interviews with ten researchers with technical backgrounds who focus on human state modeling using longitudinal passive mobile sensing. Our interview results reflect two important aspects of the state-of-the-art research: (1) common practices when researchers conduct their projects, and (2) researchers' challenges and needs during the stages of data collection, data cleaning, and data analysis. We also identified essential roles that researchers play in the research process, that is, roles that they would not want to give up, as well identifying the uniqueness of mobile sensing data for data collection and analysis. Our findings provide practical suggestions for future researchers and practitioners, and high-level design guidelines for a toolkit that could facilitate the whole research process. We envision our findings benefiting a wide range of audiences, but are especially for researchers who are interested in but do not have much previous experience in mobile passive sensing, and intermediate researchers who are working to improve their skills in the area.

Declarations

Financial interests The authors declare they have no financial interests.

Non-financial interests Author Anind K. Dey is one of the editors-in-chief of "CCF Transactions on Pervasive Computing and Interaction".

References

Aledavood, T., Torous, J., Hoyos, A.M.T., Naslund, J.A., Onnela, J.P., Keshavan, M.: Smartphone-based tracking of sleep in depression, anxiety, and psychotic disorders. Curr. Psychiatry Rep. **21**(7), 49 (2019)

Alexander, B., Karakas, K., Kohout, C., Sakarya, H., Singh, N., Stachtiaris, J., Barnes, L.E., Gerber, M.S.: A behavioral sensing system that promotes positive lifestyle changes and improves metabolic control among adults with type 2 diabetes. In: 2017 Systems and Information Engineering Design Symposium (SIEDS), IEEE, pp 283–288 (2017)

Althoff, T., Hicks, J.L., King, A.C., Delp, S.L., Leskovec, J., et al.: Large-scale physical activity data reveal worldwide activity inequality. Nature **547**(7663), 336–339 (2017)

Bae, S., Ferreira, D., Suffoletto, B., Puyana, J.C., Kurtz, R., Chung, T., Dey, A.K.: Detecting drinking episodes in young adults using smartphone-based sensors. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(2), 5 (2017)

Bae, S., Chung, T., Ferreira, D., Dey, A.K., Suffoletto, B.: Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. Addict. behav. **83**, 42–47 (2018)

Banovic, N., Buzali, T., Chevalier, F., Mankoff, J., Dey, A.K.: Modeling and understanding human routine behavior. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, pp. 248–260 (2016)

Ben-Zeev, D., Wang, R., Abdullah, S., Brian, R., Scherer, E.A., Mistler, L.A., Hauser, M., Kane, J.M., Campbell, A., Choudhury, T.: Mobile behavioral sensing for outpatients and inpatients with schizophrenia. Psychiatr. Serv. **67**(5), 558–561 (2016)

Biel, J.I., Martin, N., Labbe, D., Gatica-Perez, D.: Bites'n'Bits: inferring eating behavior from contextual mobile data. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(4), 1–33 (2018). https://doi.org/10.1145/3161161

Bilogur, A.: Missingno: a missing data visualization suite. J. Open Source Softw. **3**(22), 547 (2018)

Blunck, H., Bouvin, N.O., Franke, T., Grønbæk, K., Kjaergaard, M.B., Lukowicz, P., Wüstenberg, M.: On heterogeneity in mobile sensing applications aiming at representative data collection. In: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, pp. 1087–1098 (2013)

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: Towards crime prediction from demographics and mobile data. In: Proceedings of the International Conference on Multimodal Interaction, pp. 427–434 (2014). https://doi.org/10.1145/2663204.2663254

Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., Pentland, A.: Moves on the street: classifying crime hotspots using aggregated anonymized data on people dynamics. Big data **3**(3), 148–158 (2015)

Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. (CSUR) **46**(3), 1–33 (2014)

Canzian, L., Musolesi, M.: Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, pp. 1293–1304 (2015)

Carreiro, S., Fang, H., Zhang, J., Wittbold, K., Weng, S., Mullins, R., Smelson, D., Boyer, E.W.: imstrong: deployment of a biosensor system to detect cocaine use. J. Med. Syst. **39**(12), 186 (2015)

Carreiro, S., Wittbold, K., Indic, P., Fang, H., Zhang, J., Boyer, E.W.: Wearable biosensors to detect physiologic change during opioid use. J. Med. Toxicol. **12**(3), 255–262 (2016)

Carter, S., Mankoff, J., Klemmer, S.R., Matthews, T.: Exiting the cleanroom: on ecological validity and ubiquitous computing. Hum. Comput. Interact. **23**(1), 47–99 (2008)

Centellegher, S., Miritello, G., Villatoro, D., Parameshwar, D., Lepri, B., Oliver, N.: Mobile money: understanding and predicting its adoption and use in a developing economy. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2(4) (2018). http://arxiv.org/abs/1812.03289

Chang, L., Lu, J., Wang, J.U., Chen, X., Fang, D., Tang, Z.O., Wang, J., Tang, Z., Nurmi, P., Wang, Z.: SleepGuard: capturing rich sleep information using smartwatch sensing data. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(98), 34 (2018). https://doi.org/10.1145/3264908

Chen, X., Wang, Y., He, J., Pan, S., Li, Y., Zhang, P.: CAP?: context-aware app usage prediction with heterogeneous graph embedding. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **3**(1), 4 (2019)

Chintha, K.K., Indic, P., Chapman, B., Boyer, E.W., Carreiro, S.: Wearable biosensors to evaluate recurrent opioid toxicity after naloxone administration: a hilbert transform approach. In: Proceedings of the Annual Hawaii International Conference on System Sciences. Annual Hawaii International Conference on System Sciences, NIH Public Access, vol. 2018, p. 3247 (2018)

Cho, H., Ippolito, D., Yu, Y.W.: Contact tracing mobile apps for covid-19: privacy considerations and related trade-offs (2020). arXiv preprint arXiv:200311511

Choudhury, T., Borriello, G., Consolvo, S., Haehnel, D., Harrison, B., Hemingway, B., Hightower, J., Koscher, K., LaMarca, A., Landay, J.A., et al.: The mobile sensing platform: an embedded activity recognition system. IEEE Pervas. Comput. **7**(2), 32–41 (2008)

Deslandes, A., Moraes, H., Ferreira, C., Veiga, H., Silveira, H., Mouta, R., Pompeu, F.A., Coutinho, E.S.F., Laks, J.: Exercise and mental health: many reasons to move. Neuropsychobiology **59**(4), 191–198 (2009)

Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., Vaitla, B., González, M.C.: Sequence of purchases in credit card data reveal life styles in urban populations. Nat. Commun. **2018**, 1–8 (2017). https://doi.org/10.1038/s41467-018-05690-8. arXiv:1703.00409

Doryab, A., Dey, A.K., Kao, G., Low, C.: Modeling biobehavioral rhythms with passive sensing in the wild: a case study to predict readmission risk after pancreatic surgery. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **3**(1), 8 (2019a)

Doryab, A., Villalba, D.K., Chikersal, P., Dutcher, J.M., Tumminia, M., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J.D., et al.: Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. JMIR mHealth uHealth **7**(7), e13209 (2019b)

Fan, J., Fan, X., Tian, F., Li, Y., Liu, Z., Sun, W., Wang, H.: What is that in your hand? recognizing grasped objects via forearm electromyography sensing. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(4), 1–24 (2018)

Ferreira, E., Ferreira, D.: Towards altruistic data quality assessment for mobile sensing. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, pp. 464–469 (2017)

Ferreira, D., Kostakos, V., Dey, A.K.: Aware: mobile context instrumentation framework. Front. ICT **2**, 6 (2015)

Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009)

Harari, G.M., Müller, S.R., Aung, M.S., Rentfrow, P.J.: Smartphone sensing methods for studying behavior in everyday life. Curr. Opin. Behav. Sci. **18**, 83–90 (2017)

Hernández, N., Castro, L.A., Favela, J., Michán, L., Arnrich, B.: Data quality in mobile sensing datasets for pervasive healthcare. In: Handbook of Large-Scale Distributed Computing in Smart Healthcare. Springer, Berlin, pp. 217–238 (2017)

Imai, R., Tsubouchi, K., Konishi, T., Shimosaka, M.: Early destination prediction with spatio-temporal user behavior patterns. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(4), 1–19 (2018). https://doi.org/10.1145/3161197

Klemmer, S.R., Li, J., Lin, J., Landay, J.A.: Papier-mache: Toolkit support for tangible input. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, CHI '04, p 399–406. (2004). https://doi.org/10.1145/985692.985743

Lane, N.D., Xu, Y., Lu, H., Hu, S., Choudhury, T., Campbell, AT., Zhao, F.: Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 355–364 (2011)

Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. IEEE Commun. Mag. **48**(9), 140–150 (2010)

Laport-López, F., Serrano, E., Bajo, J., Campbell, A.T.: A review of mobile sensing systems, applications, and opportunities. Knowl. Inf. Syst. **62**(1), 145–174 (2020)

Lee, C., Lee, G.G.: Information gain and divergence-based feature selection for machine learning-based text categorization. Inf. Process. Manag. **42**(1), 155–165 (2006)

Lu, J., Bi, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., Bamis, A., Russell, A., Wang, B.: Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(1), 1–21 (2018). https://doi.org/10.1145/3191753. arXiv:1505.02818v1

Madan, A., Cebrian, M., Moturu, S., Farrahi, K., Pentland, A.: Sensing the health state of a community. IEEE Pervas. Comput. **11**(4), 36–45 (2012). https://doi.org/10.1109/MPRV.2011.79

Manek, A.S., Shenoy, P.D., Mohan, M.C., Venugopal, K.: Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and svm classifier. WWW **20**(2), 135–154 (2017)

Matarazzo, T.J., Pakzad, S.N.: Structural identification for mobile sensing with missing observations. J. Eng. Mech. **142**(5), 04016021 (2016)

Mazilu, S., Blanke, U., Calatroni, A., Gazit, E., Hausdorff, J.M., Tröster, G.: The role of wrist-mounted inertial sensors in detecting gait freeze episodes in Parkinson's disease. Pervas. Mob. Comput. **33**, 1–16 (2016)

Mehrotra, A., Tsapeli, F., Hendley, R., Musolesi, M.: Mytraces: investigating correlation and causation between users' emotional states and mobile phone interaction. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(3), 83 (2017)

Migrating apps to android 9: android developers. (2019). https://developer.android.com/about/versions/pie/android-9.0-migration

Min, J.K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.I.: Toss'n'turn: smartphone as sleep and sleep quality detector. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 477–486 (2014)

Min, C., Lee, S., Lee, C., Lee, Y., Kang, S., Choi, S., Kim, W., Song, J.: Pada: Power-aware development assistant for mobile sensing applications. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Association for Computing Machinery, New York, NY, USA, UbiComp '16, pp. 946–957. (2016). https://doi.org/10.1145/2971648.2971676

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P., Campbell, A.T., Chawla, N.V., Swain, V.D., Choudhury, M.D., Dey, A.K., D'Mello, S.K., Gao, G., Gregg, J.M., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G., Martinez, G.J., Mattingly, S.M., Moskal, E., Mulukutla, R., Nepal, S., Nies, K., Reddy, M.D., Robles-Granda, P., Saha, K., Sirigiri, A., Striegel, A.: Differentiating higher and lower job performers in the workplace using mobile sensing. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **3**(2), 1–24 (2019). https://doi.org/10.1145/3328908

Morgan, W.P., Goldston, S.E.: Exercise and Mental Health. Taylor & Francis, US (1987)

Naughton, F., Hopewell, S., Lathia, N., Schalbroeck, R., Brown, C., Mascolo, C., McEwen, A., Sutton, S.: A context-sensing mobile phone app (q sense) for smoking cessation: a mixed-methods study. JMIR mHealth uHealth **4**(3), e106 (2016)

Obuchi, M., Huckins, J.F., Wang, W., daSilva, A., Rogers, C., Murphy, E., Hedlund, E., Holtzheimer, P., Mirjafari, S., Campbell, A.: Predicting brain functional connectivity using mobile sensing. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **4**(1), 1–22 (2020)

Oliver, N., Letouzé, E., Sterly, H., Delataille, S., De Nadai, M., Lepri, B., Lambiotte, R., Benjamins, R., Cattuto, C., Colizza, V., et al.: (2020) Mobile phone data and covid-19: missing an opportunity? arXiv preprint arXiv:200312347

Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M., Pedersen, L., Petersen, I.: Missing data and multiple imputation in clinical epidemiological research. Clin. Epidemiol. **9**, 157 (2017)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

Pérez-Torres, R., Torres-Huitzil, C., Galeana-Zapién, H.: Power management techniques in smartphone-based mobility sensing systems: a survey. Pervas. Mob. Comput. **31**, 1–21 (2016)

Postolache, G., Postolache, O.: Smartphone sensing technologies for tailored parkinson's disease diagnosis and monitoring. In: Mobile Solutions and Their Usefulness in Everyday Life. Springer, Berlin, pp. 251–273 (2019)

Qin, T., Shangguan, W., Song, G., Tang, J.: Spatio-temporal routine mining on mobile phone data. ACM TKDD **12**(5), 56 (2018)

Reis, H.T.: Why researchers should think" real-world": a conceptual rationale (2012)

Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., Mohr, D.C.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J. Med. Internet Res. **17**(7), e175 (2015)

Salekin, A., Eberle, J.W., Glenn, J.J., Teachman, B.A., Stankovic, J.A.: A weakly supervised learning framework for detecting social anxiety and depression. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(2), 26 (2018)

Sarda, A., Munuswamy, S., Sarda, S., Subramanian, V.: Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study. JMIR mHealth uHealth **7**(1), e11041 (2019)

Shoaib, M., Bosch, S., Scholten, H., Havinga, P.J., Incel, O.D.: Towards detection of bad habits by fusing smartphone and smartwatch sensors. In: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), IEEE, pp. 591–596 (2015)

Srinivasan, V., Koehler, C., Jin, H.: RuleSelector: selecting conditional action rules from user behavior patterns. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(1), 1–34 (2018). https://doi.org/10.1145/3191767

Suhara, Y., Xu, Y., Pentland, A.: Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 715–724 (2017)

Sun, X., Qiu, L., Wu, Y., Tang, Y., Cao, G.: SleepMonitor: monitoring respiratory rate and body position during sleep using smartwatch. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(3), 1–22 (2017). https://doi.org/10.1145/3130969

Trifan, A., Oliveira, M., Oliveira, J.L.: Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. JMIR mHealth uHealth **7**(8), e12649 (2019)

Vaizman, Y., Weibel, N., Lanckriet, G.: Context recognition in-the-wild: unified model for multi-modal sensors and multi-label classification. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(4), 1–22 (2017). https://doi.org/10.1145/3161192

Valero-Mora, P., Rodrigo, M.F., Sanchez, M., SanMartin, J.: A plot for the visualization of missing value patterns in multivariate data. Pract. Assess. Res. Eval. **24**(1), 9 (2019)

Wang R., Aung, M.S., Abdullah, S., Brian, R., Campbell, A.T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E.A., et al.: Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 886–897 (2016b)

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, pp. 3–14 (2014)

Wang, W., HARARI, G.M., Wang, R., Muller, S.R., Mirjafari, S., Masaba, K., Campbell, A.T.: Sensing behavioral change over time : using within-person variability features from mobile sensing to predict personality traits. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **2**(3) (2018)

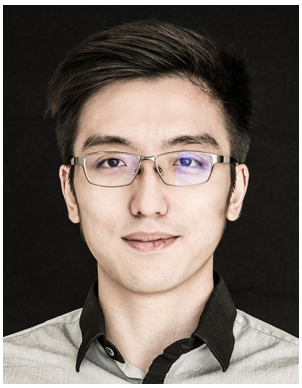
Wang, R., Harari, G., Hao, P., Zhou, X., Campbell, A.T.: Smartgpa: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, pp. 295–306 (2015)

Wang, D., Weisz, J.D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., Gray, A.: Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. Proc. ACM Hum. Comput. Interact. 3(CSCW):1–24 (2019a)

Wang, L., Zhang, D., Wang, Y., Chen, C., Han, X., M'hamed, A.: Sparse mobile crowdsensing: challenges and opportunities. IEEE Commun. Mag. **54**(7), 161–167 (2016a)

Wang, R., Wang, W., Aung, M.S., Ben-Zeev, D., Brian, R., Campbell, A.T., Choudhury, T., Hauser, M., Kane, J., Scherer, E.A., et al.: Predicting symptom trajectories of schizophrenia using mobile

sensing. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(3), 110 (2017)
Wang, H., Li, Y., Zeng, S., Wang, G., Zhang, P., Hui, P., Jin, D.: Modeling spatio-temporal app usage for a large user population. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **3**(1), 1–23 (2019b). https://doi.org/10.1145/3314414
Xu, X., Chikersal, P., Doryab, A., Villalba, D.K., Dutcher, J.M., Tumminia, M.J., Althoff, T., Cohen, S., Creswell, K.G., Creswell, J.D., et al.: Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **3**(3), 1–33 (2019)
Yan, Z., Subbaraju, V., Chakraborty, D., Misra, A., Aberer, K.: Energy-efficient continuous activity recognition on mobile phones: an activity-adaptive approach. In: 2012 16th International Symposium on Wearable Computers, Ieee, pp. 17–24 (2012)
Zhang, Z.: Missing data imputation: focusing on single imputation. Ann. Transl. Med. **4**(1) (2016)
Zhang, X., Li, W., Chen, X., Lu, S.: Moodexplorer: towards compound emotion detection via smartphone sensing. Proc. ACM Interact. Mob. Wear. Ubiq. Technol. **1**(4), 176 (2018)
Zhou, Y., De, S., Wang, W., Wang, R., Moessner, K.: Missing data estimation in mobile sensing environments. IEEE Access **6**, 69869–69882 (2018)



Xuhai Xu is a third-year PhD student in the Information School at the University of Washington, Seattle. He earned his Bachelor's degrees in Industrial Engineering (major) and Computer Science (minor) from Tsinghua University in 2018. His research interests include human-computer interaction (HCI), ubiquitous computing, and applied machine learning.

Jennifer Mankoff is the Richard E. Ladner Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her research is focused on giving people the voice, tools and agency to advocate for themselves. She strives to bring both structural and personal perspectives to her work. For example, her recent work in the intersection of mental health and discrimination uses sensed data to explore how external risks and pressures interact with people's responses to challenging moments. Similarly, her work in fabrication of accessible technologies considers not only innovative tools that can enable individual makers but also the larger clinical and sociological challenges to disseminating and sharing designs. She received her PhD at Georgia Tech, advised by Gregory Abowd and Scott Hudson, and her B.A. from Oberlin College. Her previous faculty positions include UC Berkeley's EECS department and Carnegie Mellon's HCI Institute. She has been recognized with an Alfred P. Sloan Fellowship, IBM Faculty Fellowship and Best Paper awards from ASSETS, CHI and Mobile HCI. Some supporters of her research include Autodesk, Google Inc., the Intel Corporation, IBM, Hewlett-Packard, Microsoft Corporation and the National Science Foundation.

Anind K. Dey is a Professor and Dean of the Information School at the University of Washington. Anind is renowned for his early work in context-aware computing, an important theme in modern computing, where computational processes are aware of the context in which they operate and can adapt appropriately to that context. His research is at the intersection of human-computer interaction, machine learning, and ubiquitous computing. For the past few years, Anind has focused on passively collecting large amounts of data about how people interact with their phones and the objects around them, to use for producing detection and classification models for human behaviors of interest. He applies a human-centered and problem-based approach through a collaboration with an amazing collection of domain experts in areas of substance abuse (alcohol, marijuana, opioids), mental health, driving and transportation needs, smart spaces, sustainability, and education. Anind was inducted into the ACM SIGCHI Academy for his significant contributions to the field of human-computer interaction in 2015. Before starting at the University of Washington in 2018, Anind was the Charles M. Geschke Professor and Director of the Human-Computer Interaction Institute at Carnegie Mellon University for 4 years, and was a member of the faculty for 13 years. Previously, he was a Senior Researcher at Intel Research and an Adjunct Assistant Professor of Computer Science at UC Berkeley. Anind received his PhD and MS in computer science, and an MS in aerospace engineering from Georgia Tech, and a Bachelors in Computer Engineering from Simon Fraser University.