# Learning-Assisted Optimization in Mobile Crowd Sensing: A Survey

Jiangtao Wang, Yasha Wang, Daqing Zhang, Jorge Goncalves, Denzil Ferreira, Aku Visuri, Sen Ma

**Abstract**—Mobile Crowd Sensing (MCS) is a relatively new paradigm for collecting real-time and location-dependent urban sensing data. Given its applications, it is crucial to optimize the MCS process with the objective of maximizing the sensing quality and minimizing the sensing cost. While earlier studies mainly tackle this issue by designing different combinatorial optimization algorithms, there is a new trend to further optimize MCS by integrating learning techniques to extract knowledge, such as participants' behavioral patterns or sensing data correlation. In this article, we perform an extensive literature review of learning-assisted optimization approaches in MCS. Specifically, from the perspective of the participant and the task, we organize the existing work into a conceptual framework, present different learning and optimization methods, and describe their evaluation. Furthermore, we discuss how different techniques can be combined to form a complete solution. In the end, we point out existing limitations which can inform and guide future research directions.

**Index Terms**—Mobile Crowd Sensing, Learning, Optimization.

## 1 Introduction

Coined by Howe and Robinson in [1], the idea of crowdsourcing has become an emerging distributed problem-solving paradigm by combining the power of both human computation and machine intelligence. Furthermore, the prevalence of mobile devices and the increasing smart sensing requirements in the city have led to an alternative or complementary approach for urban sensing, called Mobile Crowd Sensing (MCS) [2], [5]. MCS leverages the inherent mobility of mobile users (i.e., participants or workers), the sensors embedded in mobile phones and the existing communication infrastructures (Wi-Fi, 4G/5G networks) to collect and transfer urban sensing data. Compared to wireless sensor networks (WSN), which are based on specialized sensing infrastructures, MCS is less costly and can obtain a higher spatial-temporal coverage.

However, every coin has its two sides. Although with the above advantages and various MCS-enabled innovative applications [8], [9], [10], [11], [12], [13], the new sensing paradigm also encounters new challenges as "humans" act as sensors [14]. First, the sensing quality problem is more complex in MCS, because human sensors are quite complex and several human factors have to be taken into account. For example, it is uncertain to predict if the participants would accept the recommended sensing tasks or not. Even if they accept the task, factors such as reliability, user preference, expertise, and mobility pattern may significantly affect how they will complete these tasks (e.g., coverage and sensing quality). Second, participating in an MCS campaign incurs extra cost (e.g., energy consumption and data transferring cost) and concerns (e.g., location privacy leak) to the participants. Keeping the cost as low as possible is beneficial for motivating participants to contribute their data. In summary, with the objective of maximizing sensing quality and minimizing sensing cost control, it is crucial to optimize the entire lifecycle of MCS, and the number of relevant research works has continuously increased in recent years.

Earlier studies mainly tackle this issue from the perspective of designing different combinatorial optimization algorithms in participant selection or task assignment. With the rapid technical progress in learning-based artificial intelligence, we notice that it is now an emerging trend to integrate the learning techniques into the research problem of MCS optimization. On the one hand, a group of studies, such as [21], [22], [23], [25], [26], [27], focus on how to understand participation behavior, and then exploit the obtained knowledge to future optimize the MCS process (such as participant selection and task assignment). On the other hand, another category of works, such as [42], [43], [44], [48], [49], [50], leverage the correlation among sensing data (such as spatial-temporal correlation) or data inference techniques to optimize MCS in several aspects, such as reducing the cost in sensing data sampling and discovering truth through sensing results aggregation.

In recent years, there are several survey or tutorial papers in the MCS research community. Some [2], [4], [5] focus on the description of overall and general picture (e.g., lifecycles, research issues, and challenges) in MCS, and others such as [15], [16], [17], [18], [19] dive into specific research topics in MCS, including incentive mechanisms [15], [16], privacy preservation [17], [18], and energy saving [19]. However, to the best of our knowledge, **there are no survey or tutorial papers summarizing how learning techniques are explored to assist the MCS optimization process.** Therefore, this motivates the need for a compre-

- *Jiangtao Wang and Daqing Zhang are with school of EECS, Peking University. Yasha Wang and Sen Ma are with National Research & Engineering Center of Software Engineering in Peking University. Jiangtao Wang, Yasha Wang, Daqing Zhang and Sen Ma are also with Key Laboratory of High Confidence Software Technologies, Ministry of Education. Jorge Goncalves is with School of Computing and Information Systems, University of Melbourne, Australia. Denzil Ferreira and Aku Visuri are with University of Oulu, Finland.*
- *Jiangtao Wang and Yasha Wang are the corresponding authors.*
- *This work was mainly supported by NSFC Grant (No. 61872010).*

hensive survey.

With the above motivation, we conduct a comprehensive survey of all publications related to learning-assisted MCS optimization via a paper selection process guided by a suggestion made in [3]. The main criteria for including a paper are: a) whether it describes a research problem in MCS or similar concepts (e.g., participatory sensing, mobile crowdsourcing, and spatial crowdsourcing), and b) does the article utilize learning techniques to optimize a certain aspect of MCS. We performed three types of literature searches before Nov 2017: a) Online digital libraries including ACM, IEEE Xplore, Springer Link, Wiley, Elsevier ScienceDirect, and Google Scholar. b) Main conference proceedings and journals in fields such as ubiquitous computing, mobile computing, and wireless sensor networks from January 2008 to Nov 2017. The specific conference proceedings and journals are on the top proceeding lists within the fields [67]. c) By searching the citations from included papers, we further discovered some additional relevant papers.

The contributions of this survey paper include:

1) We present a comprehensive survey of the literature using learning techniques to optimize the process of MCS, which is a hot topic in MCS research community but lacks survey or tutorial papers. To the best of our knowledge, this article is the first work summarizing MCS optimization techniques from a learning-assisted perspective.

2) We classify the relevant works from the perspective of both participants and tasks, with the objective of maximizing quality or minimizing cost. In addition to presenting each individual technique, we discuss how they are evaluated, analyze their relationships, and discuss how they can be combined to optimize MCS systems collaboratively.

3) We highlight the existing gaps for the state-of-the-art learning-assisted MCS optimization approaches and present some future research opportunities.

## 2 MCS and Its Optimization

In this section, we present some basic background knowledge about MCS and its optimization. For more detailed understanding about MCS and its main research issues, interested readers can refer to other surveys and tutorials [2], [4], [5].

### 2.1 Preliminary of MCS

Compared to general crowdsourcing, MCS have two unique features. (1) Mobility-Relevant Features. Different from general crowdsourcing tasks, MCS requires the workers to complete sensing tasks in certain locations, because the sensing results are location-dependent (e.g., air quality, noise level, and traffic congestion status). (2) Sensing-Relevant features. Different from general crowdsourcing, MCS always targets at urban sensing tasks. First, the execution of sensors and localization modules introduces much more energy consumption into MCS than general crowdsourcing. Therefore, it is important to control the energy consumption of workers in the MCS systems. Second, many MCS tasks need to invoke phone-embedded sensors for task completion, but the set of sensors for each worker may be different as they hold various brands and models of smart devices.

Similar to the notion of participatory sensing [6] and human-centric computing [7], there are two key players in MCS, i.e., *participants* who collect and report sensing data through a mobile device, and *task organizers* who manage and coordinate the whole MCS process. The life-cycle of MCS can be divided into four stages: task creation, task assignment, task execution and data aggregation. The main functionality and research issues of each stage are briefly described as follows:

a) *Task Creation*: The MCS organizer creates an MCS task to be given to workers with the corresponding mobile applications. In this stage, the key research issue is how to reduce the time and the technical threshold of task creation [62], [63].

b) *Task Assignment*: After the organizer creates an MCS task, the next stage is task assignment, in which the MCS platform selects participants and assigns them with the different sensing tasks. The key research issue at this stage is how to optimize MCS taking into account a number of different factors, such as spatial coverage, incentive cost, energy consumption, and task completion time [29], [64].

c) *Task Execution*: Once the participants have received the assigned micro-sensing tasks, they can complete them within a pre-defined spatial-temporal scale (*i.e.*, time duration and target region). This state includes sensing, computing, and data uploading. How to save energy consumption and protect users' location and overall privacy are the core research challenges at this stage [18], [19].

d) *Data Integration*: This stage fuses the reported data from the crowd according to the requirements of task organizers. The key issue at this stage is how to infer missing data and provide a complete spatial-temporal picture of the target phenomenon (*e.g.*, real-time air quality map of a city) [43], [45].

### 2.2 Two Aspects of MCS Optimization: Quality and Cost

For the optimization of MCS, the control of sensing quality and cost is a fundamental research problem. On the one hand, we want to maximize the sensing quality of an MCS task. The sensing quality metric can be diverse for different applications (*e.g.*, spatial-temporal coverage, Quality-of-Information, mean error rate, etc.) [20]. On the other hand, we need to control the cost during the MCS process. The cost may include incentive rewards, energy consumption, data transferring expense, privacy leak, attention occupation, etc. [15], [16], [17], [18], [19].

However, the sensing quality maximization and sensing cost minimization are usually two opposing objectives. For example, to optimize the spatial-temporal coverage, we may need to recruit more participants, which will then lead to a higher total sensing cost. Therefore, how to achieve a good tradeoff between sensing quality and cost is a major research issue in MCS.

## 3 Learning-based MCS Optimization: A Conceptual Framework

In Fig 1, we present a conceptual framework for learning-based MCS optimization, which primarily consists of the following two phases.

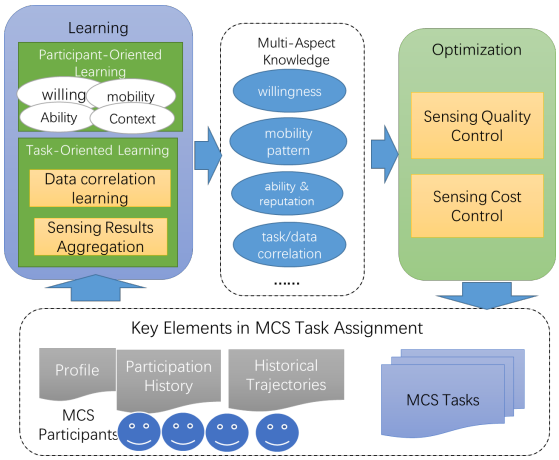



Fig. 1: Conceptual Framework for Learning-Based MCS Optimization

***Learning Phase***: we can extract knowledge from both participants and tasks. In terms of participants, with various machine learning techniques (such as classification, clustering, and regression), we can form a better understanding towards both the individual or the community of the participants for several aspects, such as willingness, mobility pattern, sensing context, ability, and reputation. In terms of tasks, it analyzes and discovers the correlation between different types of sensing data and tasks.

***Optimization Phase***: we can leverage the extracted knowledge to optimize the MCS campaign in the following aspects:

1) *Sensing quality control*. MCS faces the challenge of low-quality or even erroneous data collection. For example, a smartphone may report inaccurate data samples when it is located in a bag or pocket [21], or a participant may report malicious sensing data for his own benefit [2]. With this in mind, we need to integrate the extracted knowledge to enable quality-optimized MCS. For example, in both the participant selection and ground truth inference phase, we should assign a higher priority to the participants who are more willing to accept tasks, more reliable, and with better spatial-temporal coverage.

2) *Sensing cost control*. The process of participation in MCS campaigns leads to costs such as energy consumption, data transferring fee, and attention occupation for the participants. To compensate these, the task organizer needs to pay incentive rewards to motivate a large number of participants. The extracted knowledge in the learning phase can help us reduce cost. For example, with spatial correlation in mind, we can select a subset of more informative areas (i.e., having the highest information gain in terms of deducing the sensing data in other unselected areas), and then deduce the sensing data in unselected ones.

The above process is iterative in nature, in which the behavior data about participants and sensing data of MCS tasks are collected continuously to update the multi-aspect knowledge. Then, the updated knowledge will be further used in the MCS process.

## 4 Learning and Optimization Techniques

In this section, we present the existing approaches for optimizing MCS through learning techniques and summarize their contributions. We divide the state-of-the-art studies into the following groups: *(1) Participant-Oriented Learning*: based on the participants' profile, historical mobility traces, and participation records, we can learn and predict participants' behavior in MCS, which can be leveraged to recruit and select more beneficial participants, or assist them to better complete sensing tasks. *(2) Task-Oriented Learning*: the objective of this group of research works is to mine the data correlation in MCS tasks, and then exploit this to reduce the sensing cost or improve sensing quality.

### 4.1 Participant-Oriented Learning

A number of research studies use a data-driven approach to learn participants' behavioral patterns and exploit it in assigning tasks to more preferred participants. As a given study may involve several aspects, Table 1 summarizes this set of works in terms of the learned knowledge.

1) *Willingness*. Most of existing works (such as [29], [30], [31], [32], [33]) assume that once a participant is assigned with a task, she/he will accept and complete it. However, this is not true in real-world settings, as participants may reject the task due to several reasons. Neglecting this issue has negative impact on the performance of MCS applications. To address this problem, the authors in [22] conducted a 4-week extensive smartphone user study to explore what are the factors influencing participants' participation willingness. Their findings show that data was shared significantly more when anonymously collected, and that the data type is also an important factor. The authors in [28] carried out a study in Chicago to explore the geographic factors influencing the participation willingness, and quantitative modeling shows that travel distance to the location of the task and the socioeconomic status (SES) (i.e. a measure of ones' economic and social position based on income, education, and occupation) of the task area are important factors. These results indicate that low-SES areas are currently less able to take advantage of the benefits of MCS. In a mobile crowdsourcing framework named GP-Selector [23], the authors developed a multi-classifier based approach to infer if a participant will accept an MCS task or not, where the influencing features are the incentive reward, domain interest, task workload, and privacy concern. In the focused scenario of [24], the authors assume that the participants decide whether to accept the task based on the incentive reward and movement distance. They developed a SVM-based method to learn the relationship between task acceptance rate and these two factors, and then utilize it to design better pricing mechanisms, with the objective of reducing sensing cost while ensuring task completion. In [25], the authors have taken participants' rejection into consideration and tried to maximize the overall acceptance in order to improve the system throughput. Lastly, whether a person can be interrupted in a given situation also influences the likelihood of willingness, as explored in [68], especially if the contribution relies on manual reporting.

2) *Mobility Pattern*. Contrary to generic online crowdsourcing, MCS requires the participants' physical movement

to specific locations for task completion. Thus, the mobility pattern of the participants significantly affects the task assignment process. In [26], [27], based on a real-world deployed MCS platform in campus, authors provided an analysis for the efficiency of recommending tasks based on predicted movement patterns of individual workers. With the goal of optimizing the spatial-temporal coverage in budget-constrained MCS, a group of works such as [29], [30], [31], [32], [33] studied the optimal task allocation based on the learning participants' mobility pattern from the previous trajectories. For example, [29], [31], [32], [33] assumed that the number of calls in each spatial-temporal cell follows a Poisson distribution, and they calculate the probability of participants' presence in each spatial-temporal cell based on historical trajectories. The authors in [30] adopted a location probability transition approach (i.e., calculating the transition probability between two locations) to accomplish mobility learning and prediction.

3) *Sensing Context*. Sensing context (e.g., the participants' motion and the position of the mobile device) has a significant impact on the sensing data quality for certain types of MCS tasks. The authors in [34] trained a sensing data quality classifier, which extract the relation between context information (such as the participants' motion) and sensing data quality, to estimate data quality in MCS. This classifier can be applied to guide user recruitment and task assignment in MCS.

4) *Ability and Reputation*. Learning participants' abilities and reputations can help selecting more capable and reliable participants [35], [36], [37], [38], [39]. For instance, through an empirical study, [35] revealed that participants' cognitive abilities correlate tightly with their crowdsourcing performance, where they built two models for crowdsourcing task performance prediction. In another example, [36] proposed a reputation-based system that employs the Gompertz function for learning the participants' reputation score, and implement this idea in the scenario of a crowd noise level monitoring application. Though with different definitions of reputation metrics, they learn the reputation scores in either of the two categories: 1) statistical reputation scores that are computed based on the comparison between reported data and estimated the ground truth. 2) vote-based reputation scores by the participants of MCS.

### 4.2 Task-Oriented Learning Approaches

Learning techniques also can be used to extract knowledge from the perspective of the tasks. Here we will present how the learning approach can optimize MCS in sensing data correlation learning and sensing data aggregation.

#### 4.2.1 Sensing Data Correlation Learning

Learning and exploiting sensing data correlation is an important technique to optimize the MCS process. It is based on the notion that, typically, there is a correlation among diverse sensing targets in the real world, and we can use this to address the sensing data redundancy and sparsity issues in MCS. By appropriately using data correlation, we can require the participants to collect only a relatively small number of data samples and deduce more information, thus the cost of MCS is significantly reduced.

In recent years, a number of studies in MCS focus on these aspects. Both [40] and [41] investigated a traffic status monitoring task, in which they use the correlation between the traveling speed on different roads sections to maximize the sensing accuracy with a fixed number of crowd sensors. The authors in [42], [43], [44], [45] utilized the spatial-temporal correlation of environmental sensing data (e.g., temperature and air quality) to achieve an optimized tradeoff between sensing cost and quality, in which they use matrix completion technology to infer the missing sensing data. The study in [46] demonstrated the feasibility of applying compressive sensing to data domains like large-scale question-based user surveys. The approach proposed in [47] is the extension of [46], which considered the sensing data reliability in different subareas due to different sampling density. Both [48] and [49] built a dependency graph between different entities in the city (such as the availability of shops and gas stations) to increase fact-finding accuracy. Focusing on the scenario where MCS is utilized to collect training data of context-aware applications, [50] proposed an active learning framework for optimally budgeted MCS. The authors in [73] exploited the spatial-temporal correlation of users' mobility to achieve the tradeoff between MCS task performance and privacy preserving objective.

Although the above literature is different in terms of data type and detailed algorithms, they also attempt to address one of the three important issues: *1) Informative Sampling*: how to select the most informative data collections? *2) Missing Data Inference*: how to infer the missing data from the obtained one? *3) Quality Estimation*: how to estimate if the inference meets the accuracy requirement without ground-truth sensing data. The summary is in Table 2.

#### 4.2.2 Sensing Results Aggregation

Different from the traditional wireless sensor network, MCS faces the challenge of unreliable data samples due to many reasons (e.g., uncertain sensing context and malicious participants). To achieve high-quality results, we need to collect sensing data from multiple participants for the same sensing target and infer the truth. This problem is similar to truth discovery, which has been studied extensively in the general crowdsourcing community. Specifically, there are two inputs, i.e., the task answers and the expertise of each participant. Recently, a survey has comprehensively summarized this topic [51], where most of the literature [52], [53] use voting-based strategies, such as majority voting, weighted voting, Bayesian voting, etc.

Different from general online crowdsourcing, the truth discovery problem in MCS is more complex because of the multi-modality nature and spatial-temporal features of the sensing data, and some participant-side factors (e.g., location privacy). Thus, the techniques that existing works adopted for truth discovery in MCS are different to some degree. A number of works [54], [55], [56], [57], [58] leveraged Expectation Maximization (EM) based algorithms to estimate the reliability of participants or mobile devices, which will be used as the weight to infer the ground truth of sensing data. Some other works [59], [60] have adopted unsupervised learning approaches, in which they employ an additional optimization objective to improve the EM-based

TABLE 1: Learning participant-side factors to optimize MCS: a summary

| References | Willingness | Mobility | Sensing context | Ability | Reputation |
|---|---|---|---|---|---|
| [22] | **Yes** | | | | |
| [23] | **Yes** | **Yes** | | **Yes** | **Yes** |
| [24] | **Yes** | **Yes** | | | |
| [28] | **Yes** | **Yes** | | | |
| [29], [30], [31], [32], [33] | | **Yes** | | | |
| [34] | | **Yes** | **Yes** | | |
| [35] | | | | **Yes** | |
| [36], [37], [38], [39] | | | | | **Yes** |

TABLE 2: A summary of studies to optimize MCS through the learning of the data correlation

| Literatures | Data Type | Informative Sampling | Missing Data Inference | Quality Estimation |
|---|---|---|---|---|
| [40] | Traffic speed | Heuristic greedy | Markov random field | Not addressed |
| [41] | Traffic speed | Not addressed | Matrix completion | Not addressed |
| [42], [43] | Temperature, air quality | The variance of different inference algorithm | Matrix completion | Leave-one-out estimation |
| [44] | Air quality | Not addressed | Matrix completion | Not addressed |
| [45] | Temperature, air quality, traffic speed | The variance of different inference algorithm | Matrix completion | Leave-one-out estimation |
| [46] | Question-based user surveys | Compressive sensing | Matrix completion | Not addressed |
| [48], [49] | Availability of urban entities | Not addressed | Bayesian Network | Not addressed |
| [50] | Labels and training data for activity recognition apps | Active Learning | Not addressed | Not addressed |

method. Recently, truth discovery concerning the privacy-preserving issue has been studied [61], which infers the missing data using matrix factorization techniques.

## 5 How to Conduct Evaluation

One important question about the research on learning-based MCS optimization is that: *Where can we get the training data, and how to evaluate the performance of a given approach?* We know that the ideal way is to obtain large-scale data about participants' behavior and collected sensing data, based on which extensive evaluation can be conducted. However, it is difficult to conduct such a large-scale and real-world evaluation as platforms, such as Amazon Mechanical Turk and WAZE, are not willing to open their data due to commercial reasons. Thus, researchers adopt alternative ways to demonstrate the feasibility of their proposed approach. In this section, we summarize different methodologies, which we hope can inspire and support the evaluation of future research efforts.

By summarizing the existing work, the evaluation methodology can be divided into the following three categories.

1) *Small-scale real-world evaluation*. A group of studies develops their own testbed to collect relevant data for evaluation. For example, the authors in [26], [27] build campus-scale MCS platforms as the research testbeds, in which 80 real users are recruited to complete several types of MCS tasks within a 4-week period. Similar platforms such as gMission and ChinaCrowds are developed and utilized in studies such as [25], [56], [69], [70].

2) *Open dataset based evaluation*. Another group of research works evaluates their solutions based on an open datasets (such as D4D[1], Gowalla[2]). For example, [29], [31], [32], [33], [71] evaluate a mobility pattern learning algorithm and task assignment approach based on open data containing the mobility trace of a large number of participants (e.g., calling trace and check-in data in a social network). Furthermore, [48], [49] evaluate their dependency analysis approach with a real-world dataset about the availability of groceries, pharmacies, and gas stations during Hurricane Sandy. The authors in [42], [43] evaluate their missing data inference algorithms based on a campus-scale open dataset for temperature and air quality measures.

3) *Simulation-based evaluation*. Another alternative way is to develop a simulator, in which the agents (both the participants and task organizers) are simulated according to pre-defined rules. Then, we can use the simulated data generated by the agents to perform the evaluation. A significant number of studies adopt the simulation-based approach to evaluate their learning-based MCS optimization approaches [23], [25], [27], [29], [31], [32]. We also note that several papers published in top venues choose to conduct an evaluation of both the real-world and simulated data. This is because real-world data is always better, but they often constitute isolated points in a large space. The simulation, in contrast, can extensively test the performance under various settings. Conducting the experiments based on both these two types of data can make the research work more solid.

Actually, we believe that a promising method should be the combination of both real-world and simulative evaluation. For example, we can collect small-scale and real-world data to generate some key parameters, and use these parameters to enable a large-scale simulation. For example, in [66], the authors learn the distribution of the parameters about the participants' preferences in completing MCS tasks using real-world data from 80 participants during 4 weeks. Then, they further evaluate the proposed algorithm by a simulation study, in which the parameters are generated based on the pre-learned distribution.

1. http://www.d4d.orange.com/en/presentation/data
2. https://snap.stanford.edu/data/loc-gowalla.html

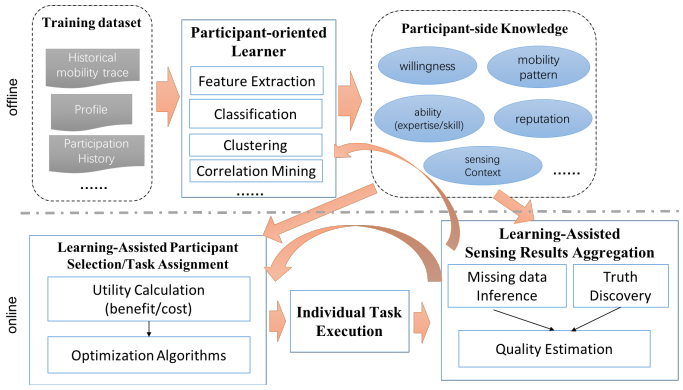



Fig. 2: Technical framework for learning-assisted MCS optimization

## 6 Combined Technical Framework

The abovementioned research works use various learning techniques to optimize MCS from a certain perspective. *Then, the interesting questions would be: what is the relationship between these works and relevant techniques? How these individual techniques can be combined together to optimize MCS in a collaborative manner?* If we can answer these two questions, it is possible that we can form a complete solution for learning-assisted MCS optimization. By summarizing and analyzing the literature on this topic, we explore the relationship between different individual research works and propose an overall technical framework shown in Fig 2.

In the offline phase, the participant-oriented learner *extracts multi-aspect knowledge about the participants*, and the output might be the classification model for predicting willingness [23], [24], [25], location [29], [30], [31], [32], [33], sensing context [34], ability and reputation [35], [36], [37], [38], [39], etc. Here we should also note that the participant-side classification models and predictors should also be retrained with updates in training data.

In the online phase, the MCS applications or platforms select participants and assign tasks based on the utility calculation, which will use the predicted participant-side factors. Intuitively, through the utility calculation, the platform may prefer to select those participants who have a higher likelihood to accept the tasks, who can obtain wider and more informative coverage in terms of mobility, and who are more reliable or capable of completing tasks. Some of the selected participants will accept and report sensing data for the assigned tasks. The server side receives the reports and will use learning techniques, such as missing data inference [42], [43], [44], [45] and truth discovery [54], [55], [56], [57], [58], to improve its quality. Then, a quality estimation method (e.g., leave-one-out) will be executed to access if the current results meet the requirement of the application [42], [43], [45]. If yes, it will generate the final output (e.g., the city-scale air quality sensing map). Otherwise, it will iteratively select participants and assign tasks until the total budget runs out.

## 7 Future Research Opportunities

In this section, we highlight the research gaps and future opportunities of learning-based MCS optimization, which may lead to novel solutions in this increasingly important field.

### 7.1 A Unified Middleware Framework

Each of the existing works tackles one specific aspect of the learning-assisted MCS optimization. To develop a real-world MCS application or system, we need to integrate different techniques to form a complete optimizing solution. In Section 6, we highlight the relationship between different techniques and present a preliminary discussion about how they may be combined to optimize the MCS process collaboratively. Thus, we argue that it is a promising research direction to study a unified learning-assisted MCS optimization framework by: 1) integrating different single techniques into the framework, and 2) exploring if we can use one to augment the others. Here, as there are multiple techniques that can be used to implement a certain component, we need to figure out what is the best combination in terms of the performance. For example, after obtaining the reports from different participants, the server may need to use missing data inference techniques to infer the data of one sensing subarea from the other (we refer this as "inter-subarea inference" in this paper). In the meantime, it also needs to discover the truth from multiple reports in the same subarea (we refer this as "intra-subarea inference" in this paper). For both inter-subarea and intra-subarea inferences, there are multiple detailed algorithms, then the challenge is to obtain the optimal combination and execution sequence. Besides, the learning-assisted MCS optimization is a common functionality across multiple applications and it is technically challenging for app developers. Thus, it is preferred that we develop a middleware framework with several application interfaces (API). Through these APIs and guidance about how these API should be combined, the technical threshold becomes lower and the developers can build their own app or system in a faster manner.

### 7.2 Leveraging Sensing Context

Mobile phones have an increasing number of sensors that can be leveraged to determine the participants' current sensing context. This includes not only hardware sensors (e.g., accelerometer, gyroscope, screen state), but also software sensors (e.g., notifications, application usage and selections). However, as shown in this survey there are not that many examples of sensing context being effectively used in MCS, particularly with regards to software sensors. Given the availability of this rich contextual information, there is a significant research opportunity to develop improved learning-assisted mechanisms for MCS optimization by leveraging this data. Hence, it is important to empirically determine the effect of different contextual factors on the likelihood of a participant completing an MCS task as well as their impact on data quality. For instance, information on how long ago a participant last used their device can provide hints on when to deliver MCS tasks. Session duration can also influence one's participation [72]. In another example, by detecting instances where a participant is bored it is then possible to take advantage of their contextual cognitive surplus [65].

### 7.3 Knowledge Transferring

Existing works mainly extract knowledge from an individual participant or task, which is an initial step to integrate learning techniques into MCS optimization. Nevertheless, this may turn out to be impractical in some real-world settings. For example, data sparsity can be a challenging issue. For new participants and tasks or due to the reason of privacy preservation, some types of historical data are not always available. In this case, studying the similarity among different participants/tasks and leveraging it in MCS optimization to tackle the data sparsity issue can be a potential research opportunity. For example, we can infer a participant's willingness and reliability from other similar ones. Alternatively, we can deduce the participant's mobility pattern in platform A through his/her traces on other platforms (such as B, C and D). It is also interesting to compare the behavioral pattern of the same set of participants on different sensing tasks, or different clusters of participants on the same sensing task, which may reveal beneficial insights about how we should design an MCS system.

### 7.4 Task Routing and Assignment

Currently, most MCS systems rely on a central authority to coordinate the task assignment process, not taking into account participants' interest and skills. There is a significant research opportunity to develop and evaluate simple and robust mechanisms to determine a participants' aptitude to complete tasks before they are assigned a sizable amount of work. For instance, a number of qualification tasks could be deployed to verify the aptitude of a participant to complete certain types of tasks. Based on their performance on those tasks, they would then either get assigned or not further tasks of the same type. This way, the amount of data collection and analysis effort could be substantially reduced.

## 8 Conclusion

In this article, we presented a survey of learning-assisted MCS optimization. Specifically, we summarized state-of-the-art research in the perspective of participant and task, and presented different learning and optimization approach together with their evaluation. Furthermore, we discussed how different individual techniques can be integrated to optimize MCS together. In the end, we highlight the gaps in this area and propose future research opportunities.

## References


[1] Howe, Jeff. "The rise of crowdsourcing." Wired magazine, 14 (6) :pp.1-4, 2006.

[2] R.K. Ganti, F. Ye, and H. Lei. Mobile crowdsensing: Current state and future challenges. IEEE Communications Magazine, 49:3239, 2011.

[3] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, Systematic literature reviews in software engineering-a systematic literature review, Information and Software Technology, vol. 51, no. 1, pp. 7 15, 2009

[4] Zhang, D., Wang, L., Xiong, H., & Guo, B. (2014). 4W1H in mobile crowd sensing. IEEE Communications Magazine, 52(8), 42-48.

[5] Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N. Y., Huang, R., & Zhou, X. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM Computing Surveys (CSUR), 48(1), 7, 2015.

[6] Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. Participatory sensing. Center for Embedded Network Sensing, 2006.

[7] Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. A., Lu, H., ... & Ahn, G. S. The rise of people-centric sensing. IEEE Internet Computing, 12(4), 2008.

[8] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, Ear-phone: an end-to-end participatory urban noise mapping system, In Proceedings of International Conference on Information Processing in Sensor Networks (IPSN 2010), pp. 105116.

[9] S. Morishita, S. Maenaka, D. Nagata, M. Tamai, K. Yasumoto, T. Fukukura, and K. Sato, Sakurasensor: quasi-realtime cherry-lined roads detection through participatory video sensing by cars, in Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing. (UBICOMP 2015), pp. 695705.

[10] Wang, J., Wang, Y., Zhang, D., Wang, L., Chen, C., Lee, J. W., & He, Y. Real-time and generic queue time estimation based on mobile crowdsensing. Frontiers of Computer Science, 11(1), 49-60, 2017.

[11] Wang, Y., Wang, J., & Zhang, X. Qtime: A queuing-time notification system based on participatory sensing data. In Proceedings of Computer Software and Applications Conference (COMPSAC, 2013), pp. 770-777.

[12] Zhou, P., Zheng, Y., & Li, M.. How long to wait?: predicting bus arrival time with mobile phone based participatory sensing. In Proceedings of 10th ACM international conference on Mobile systems, applications, and services (MobiSys 2012), pp. 379-392.

[13] Zhou T, Xiao B, Cai Z, et al. From Uncertain Photos to Certain Coverage: a Novel Photo Selection Approach to Mobile Crowdsensing, In Proceedings of IEEE International Conference on Computer Communications (INFOCOM 2018).

[14] Kostakos, V., Rogstadius, J., Ferreira, D., Hosio, S., & Goncalves, J. (2017). Human sensors. In Participatory Sensing, Opinions and Collective Awareness. Springer International Publishing, pp. 69-92, 2017.

[15] Jaimes, L. G., Vergara-Laurens, I. J., & Raij, A. A survey of incentive techniques for mobile crowd sensing. IEEE Internet of Things Journal, 2(5), 370-380, 2015.

[16] Zhang, X., Yang, Z., Sun, W., Liu, Y., Tang, S., Xing, K., & Mao, X. Incentives for mobile crowd sensing: A survey. IEEE Communications Surveys & Tutorials, 18(1), 54-67, 2016.

[17] Christin, D., Reinhardt, A., Kanhere, S. S., & Hollick, M. A survey on privacy in mobile participatory sensing applications. Journal of systems and software, 84(11), 1928-1946, 2011.

[18] Pournajaf, L., Xiong, L., Garcia-Ulloa, D. A., & Sunderam, V. A survey on privacy in mobile crowd sensing task management. Tech. Rep. TR-2014002, 2014.

[19] Jiangtao Wang, Yasha Wang, Daqing Zhang, Sumi Helal. Energy Saving Techniques in Mobile Crowd Sensing: Current State and Future Opportunities. IEEE Communications Magazine, 2018.

[20] Restuccia, F., Ghosh, N., Bhattacharjee, S., Das, S., & Melodia, T. (2017). Quality of Information in Mobile Crowdsensing: Survey and Research Challenges. ACM Transactions on Sensor Networks. 13(4), 34, 2017.

[21] Coskun, D., Incel, O. D., & Ozgovde, A. Phone position/placement detection using accelerometer: Impact on activity recognition. In Proceedings of Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2015), pp. 1-6.

[22] Gustarini, Mattia, Katarzyna Wac, and Anind K. Dey. "Anonymous smartphone data collection: factors influencing the users' acceptance in mobile crowd sensing." Personal and Ubiquitous Computing, 65-82, 2016.

[23] Wang, J., Wang, Y., Wang, L., & He, Y. GP-selector: a generic participant selection framework for mobile crowdsourcing systems. World Wide Web, 1-24, 2017.

[24] Karaliopoulos, M., Koutsopoulos, I., & Titsias, M. (2016, July). First learn then earn: Optimizing mobile crowdsensing campaigns through data-driven user profiling. In Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2016). pp. 271-280.

[25] Zheng, L., & Chen, L. Maximizing acceptance in rejection-aware spatial crowdsourcing. IEEE Transactions on Knowledge and Data Engineering, 29(9), 1943-1956, 2017.

[26] T. Kandappu, A. Misra, S.F. Cheng, N. Jaiman, R. Tandriansyah, C. Chen, H. C. Lau, D. Chander, K. Dasgupta, "Campus-Scale Mobile Crowd-Tasking: Deployment and Behavioral Insights", In Proceedings of ACM conference on Computer Supported Cooperative Work and Social Computing (CSCW 2016).

[27] T. Kandappu, N. Jaiman, R. Tandriansyah, A. Misra, S. F. Cheng, C. Chen, H. C. Lau, D. Chander, K. Dasgupta, "TASKer: Behavioral Insights via Campus-based Experimental Mobile Crowdsourcing", 2016 ACM International Joint Conference of Pervasive and Ubiquitous Computing (UbiComp 2016).

[28] Thebault-Spieker, J., Terveen, L. G., & Hecht, B. Avoiding the south side and the suburbs: The geography of mobile crowdsourcing markets. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW 2015). pp. 265-275.

[29] Zhang, D., Xiong, H., Wang, L., & Chen, G.. CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP 2014). pp. 703-714.

[30] Z. Song, C. H. Liu, J. Wu, J. Ma, and W. Wang. QoI-Aware Multitask-Oriented Dynamic Participant Selection with Budget Constraints. IEEE Transactions on Vehicular Technology, 63: 4618-4632, 2014.

[31] Wang, J., Wang, Y., Zhang, D., Xiong, H., Wang, L., & Sumi, H., et al. Fine-grained multi-task allocation for participatory sensing with a shared budget. Internet of Things Journal, 3(6), 1395-1405, 2016.

[32] Wang, J., Wang, Y., Zhang, D., Wang, F., He, Y., & Ma, L. PSAllocator: Multi-Task Allocation for Participatory Sensing with Sensing Capability Constraints. The, ACM Conference on Computer- Supported Cooperative Work and Social Computing (CSCW 2017).

[33] Liu, Y., Guo, B., Wang, Y., Wu, W., Yu, Z., & Zhang, D. TaskMe: multi-task allocation in mobile crowd sensing. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP 2016), pp. 403-414.

[34] Liu, S., Zheng, Z., Wu, F., Tang, S., & Chen, G. Context-aware data quality estimation in mobile crowdsensing. In Proceedings of IEEE Conference on Computer Communications (INFOCOM 2017), pp. 1-9.

[35] Goncalves, J., Feldman, M., Hu, S., Kostakos, V., & Bernstein, A. Task Routing and Assignment in Crowdsourcing based on Cognitive Abilities. In Proceedings of the 26th International Conference on World Wide Web (WWW 2017), pp. 1023-1031

[36] Huang, K. L., Kanhere, S. S., & Hu, W. Are you contributing trustworthy data? the case for a reputation system in participatory sensing. In Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems (MSWiM 2010), pp. 14-22

[37] Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science, 31(1), 139-167, 2017.

[38] Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., & Rampini, A. On predicting and improving the quality of Volunteer Geographic Information projects. International Journal of Digital Earth, 9(2), 134-155, 2016.

[39] Pouryazdan, M., Kantarci, B., Soyata, T., Foschini, L., & Song, H. Quantifying User Reputation Scores, Data Trustworthiness, and User Incentives in Mobile Crowd-Sensing. IEEE Access, 5, 1382-1397, 2017.

[40] Hu, H., Li, G., Bao, Z., Cui, Y., & Feng, J. Crowdsourcing-based real-time urban traffic speed estimation: From trends to speeds. In Proceedings of IEEE 32nd International Conference on Data Engineering (ICDE 2016), pp. 883-894.

[41] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A compressive sensing approach to urban traffic estimation with probe vehicles," IEEE Transactions on Mobile Computing, vol. 12, no. 11, pp. 22892302, 2013.

[42] Wang, L., Zhang, D., Wang, Y., Chen, C., Han, X., & M'hamed, A. Sparse mobile crowdsensing: challenges and opportunities. IEEE Communications Magazine, 54(7), 161-167, 2016.

[43] Wang, L., Zhang, D., Pathak, A., Chen, C., Xiong, H., Yang, D., & Wang, Y. CCS-TA: Quality-guaranteed online task allocation in compressive crowdsensing. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP 2015). pp. 683-694.

[44] Meng, C., Xiao, H., Su, L., & Cheng, Y. (2016, November). Tackling the Redundancy and Sparsity in Crowd Sensing Applications. In Proceedings of ACM Conference on Embedded Networked Sensor Systems (SenSys 2016). pp. 150-163.

[45] L. Wang, D. Zhang, D. Yang, A. Pathak, C. Chen, X. Han, H. Xiong, Y. Wang. SPACE-TA: Cost-Effective Task Allocation Exploiting Intradata and Interdata Correlations in Sparse Crowdsensing. ACM Transactions on Intelligent Systems and Technology, vol. 9, no. 2, 2017, pp. 20:1-20:28, 2017.

[46] Xu, L., Hao, X., Lane, N. D., Liu, X., & Moscibroda, T. More with less: Lowering user burden in mobile crowdsourcing through compressive sensing. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP 2015), pp. 659-670.

[47] Hao, X., Xu, L., Lane, N. D., Liu, X., & Moscibroda, T. In Proceedings of International Conference on Information Processing in Sensor Networks Density-aware compressive crowdsensing (IPSN 2017), pp. 29-39.

[48] Meng, C., Jiang, W., Li, Y., Gao, J., Su, L., Ding, H., & Cheng, Y. Truth discovery on crowd sensing of correlated entities. In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys 2015), pp. 169-182.

[49] Wang, S., Su, L., Li, S., Hu, S., Amin, T., Wang, H., ... & Abdelzaher, T. Scalable social sensing of interdependent phenomena. In Proceedings of the 14th International Conference on Information Processing in Sensor Networks (SenSys 2015), pp. 202-213.

[50] Xu, Q., & Zheng, R. When data acquisition meets data analytics: A distributed active learning framework for optimal budgeted mobile crowdsensing. In Proceedings of IEEE Conference on Computer Communications (INFOCOM 2017), pp. 1-9.

[51] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., ... & Han, J. (2016). A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 17(2), 1-16.

[52] Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., & Han, J. On the discovery of evolving truth. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015) pp. 675-684.

[53] Ma, F., Li, Y., Li, Q., Qiu, M., Gao, J., Zhi, S., ... & Han, J. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015), pp. 745-754.

[54] Wang, D., Amin, M. T., Li, S., Abdelzaher, T., Kaplan, L., Gu, S., ... & Wang, X. Using humans as sensors: an estimation-theoretic perspective. In Proceedings of 13th IEEE International Conference on Information Processing in Sensor Networks (IPSN 2014), pp. 35-46.

[55] Wang, D., Kaplan, L., Le, H., & Abdelzaher, T. On truth discovery in social sensing: A maximum likelihood estimation approach. In In Proceedings of 11th IEEE International Conference on Information Processing in Sensor Networks (IPSN 2012), pp. 233-244.

[56] Hu, H., Zheng, Y., Bao, Z., Li, G., Feng, J., & Cheng, R. Crowdsourced POI labelling: Location-aware result inference and task assignment. In Proceedings of 32nd IEEE International Conference on Data Engineering (ICDE 2016), pp. 61-72

[57] Wang, S., Wang, D., Su, L., Kaplan, L., & Abdelzaher, T. F. Towards cyber-physical systems in social spaces: The data reliability challenge. In Proceedings of Real-Time Systems Symposium (RTSS 2014), pp. 74-85

[58] Yao, S., Hu, S., Li, S., Zhao, Y., Su, L., Kaplan, L., ... & Abdelzaher, T. (2016, June). On Source Dependency Models for Reliable Social Sensing: Algorithms and Fundamental Error Bounds. In Proceedings of 36th IEEE International Conference on Distributed Computing Systems (ICDCS 2016), pp. 467-476.

[59] Yao, S., Amin, M. T., Su, L., Hu, S., Li, S., Wang, S., ... & Yener, A. Recursive ground truth estimator for social data streams. In Proceedings of 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 2016), pp. 1-12.

[60] Ouyang, R. W., Srivastava, M., Toniolo, A., & Norman, T. J. Truth discovery in crowdsourced detection of spatial events. IEEE Transactions on Knowledge and Data Engineering, 28(4), pp. 1047-1060, 2016.

[61] Miao, C., Jiang, W., Su, L., Li, Y., Guo, S., Qin, Z., ... & Ren, K. Cloud-enabled privacy-preserving truth discovery in crowd sensing systems. In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys 2015), pp. 183-196.

[62] Wang, J., Wang, Y., & Zhao, J. Helping campaign initiators create mobile crowd sensing apps: a supporting framework. In Proceedings of IEEE 39th Computer Software and Applications Conference (COMPSAC 2015), pp. 545-552.

[63] Wang, J., Wang, Y., & He, Y. Lowering the technical threshold for organizers to create and deliver mobile crowd sensing applications. International Journal of Distributed Sensor Networks, 11(11), 721647.

[64] S. Hachem, A. Pathak, and V. Issarny. Probabilistic registration for large-scale mobile participatory sensing. In Proceedings of the 2013 IEEE International conference on Pervasive Computing and Communications, volume 18, page 22, 2013.
[65] Pielot, M., Dingler, T., Pedro, J. S., & Oliver, N. When attention is not scarce-detecting boredom from mobile phone usage. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing (UBICOMP 2015), pp. 825-836.
[66] S. F. Cheng, C. Chen, T. Kandappu, H. C. Lau, A. Misra, N. Jaiman, R. Tandriansyah, D. Koh, "Scalable Urban Mobile Crowdsourcing: Handling Uncertainty in Worker Movement", ACM Transactions on Intelligent Systems and Technology, 9(3), 26, 2017.
[67] http://www.guide2research.com/topconf/
[68] A. Visuri, N. van Berkel, J. Goncalves, C. Luo, D. Ferreira, V. Kostakos. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobiHCI 2017).
[69] N. van Berkel, C. Luo, D. Ferreira, J. Goncalves and V. Kostakos. The Curse of Quantified-Self: An Endless Quest for Answers International Joint Conference on Pervasive and Ubiquitous Computing (Adjunct), 2015, 973-978.
[70] N. van Berkel, J. Goncalves, S. Hosio, V. Kostakos, "Gamification of Mobile Experience Sampling Improves Data Quality and Quantity", Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), vol. 1, no. 3, pp. 107:1-107:21, 2017.
[71] D. Ferreira, V. Kostakos and I. Schweizer. Human Sensors on the Move. In Participatory Sensing, Opinions and Collective Awareness, Springer International Publishing, 9-19, 2017
[72] A. Visuri, N. van Berkel, C. Luo, J. Goncalves, D. Ferreira, V. Kostakos, "Challenges of Quantified-Self: Encouraging Self-Reported Data Logging During Recurrent Smartphone Usage", in British HCI, 2017.
[73] Linghe Kong, Liang He, Xiao-Yang Liu, Yu Gu, Min-You Wu, Xue Liu. "Privacy-Preserving Compressive Sensing for Crowdsensing based Trajectory Recovery". IEEE ICDCS, Columbus, Ohio, USA, 2015.

**Jiangtao Wang** received his Ph.D. degree in Peking University, Beijing, China, in 2015. He is currently an assistant professor in Institute of Software, School of Electronics Engineering and Computer Science, Peking University. His research interest includes mobile crowd sensing and social computing.

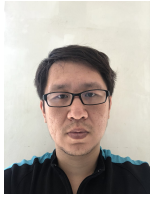
**Yasha Wang** received his Ph.D. degree in Northeastern University, Shenyang, China, in 2003. He is a professor of National Research & Engineering Center of Software Engineering in Peking University, China. His research interest includes urban data analytics and ubiquitous computing.

**Daqing Zhang** is a professor at Peking University, China, and Tlcom SudParis, France. He obtained his Ph.D from the University of Rome "La Sapienza," Italy, in 1996. His research interests include context-aware computing, urban computing, mobile computing, and so on.

**Jorge Goncalves** is a Lecturer in the School of Computing and Information Systems at the University of Melbourne, Australia. Goncalves received PhD from the University of Oulu. His research interests include ubiquitous computing, humancomputer interaction, and social computing.

**Denzil Ferreira** is an Adjunct Professor at the University of Oulu, Finland. He holds a PhD in Computer Science (2013, University of Oulu). His research interests are on sensor instrumentation, human behavior modeling and context awareness.

**Aku Visuri** is currently a PhD student at the Center of Ubiquitous Computing at University of Oulu. His research interests include ubiquitous computing and mobile computing.



**Sen Ma** received his Ph.D. degree in Peking University, Beijing, China, in 2014. He is currently an assistant professor in Software Engineering Research Center, Peking University. His research interest includes software analysis and social computing.