# Crowdsourcing for Large-Scale Pervasive Sensing

**Deepak Ganesan**, **Mark Corner**
*Department of Computer Science*
*University of Massachusetts Amherst, MA 01003*
*{dganesan,mcorner}@cs.umass.edu*

Crowdsourcing, or the act of outsourcing a task to the crowd, has the potential to revolutionize information collection and processing systems by enabling in-depth, large-scale, cost-effective information gathering, and more accurate techniques for information extraction from data. Crowdsourcing provides a powerful mechanism for creating data about the physical world, particularly through the use of mobile phones and their rich set of on-board sensors (GPS, audio, video, etc). These sensors can be utilized to provide continuous and unprecedented visibility into the state of the world across many scales. Crowdsourcing is also effective when humans are better than existing automated computer algorithms, for example labeling images, transcribing speech, annotating text, transcribing scanned documents, language translation, and others. Our research seeks to leverage crowdsourcing to address several hard challenges in pervasive computing systems.

**Crowdsourcing for Data Collection** Despite the ability to leverage app markets for smartphones, researchers still face a significant barrier to doing experiments at scale. The barriers to scaling are many: 1) competition with hundreds of thousands of applications on mobile app markets, 2) investment in time and energy to design a robust, scalable, and visually appealing application and backend infrastructure, 3) limited retention from users, who rarely use applications beyond a few weeks, 4) handling human subjects, privacy, incentives, data quality, and several other challenges that are intrinsic to the use of the crowds, and others. Thus, while the idea of scaling to millions of devices is appealing, the pervasive computing community still largely relies on expensive and short-term user studies with limited numbers of users.

mCrowd[1] is a research enabler for pervasive computing at scale, inspired by crowdsourcing systems such as the Amazon Mechanical Turk. mCrowd enables a researcher to rapidly create "tasks" involving data collection from users and sensors on the phone such as audio, video, images, surveys, GPS traces, wireless connectivity traces, and others. Users who download our application for mobile phones can participate in any of these data collection efforts, in return for a reward. mCrowd will also provide several APIs for incentives, privacy, data quality management, and sensor processing that will enable a user to leverage the framework for their specific data collection goals. By providing access to a large set of people who are willing to participate in research studies for low pay, we hope to spur research innovation in a variety of disciplines that can use data from people and phones.

Data collection from the masses opens up a spectrum of research problems. How can we incentivize sensor data collection to scale to millions of users? What is the relationship between reward and delay? How can we use the system for diverse end-applications such as healthcare, traffic monitoring, citizen journalism, and others? How can we design techniques to process and filter mobile sensor data to extract useful, actionable information? How can we filter redundant information and spam? How can we handle

[1] http://crowd.cs.umass.edu

malicious users whose only goal is to maximize their rewards? How can we target information gathering to the specific users who might have the most valuable data for a task? What are the implications on privacy? While such questions have been looked at in narrower user studies in the past, mCrowd enables researchers to explore them in the context of a real, crowdsourced marketplace at significantly larger scale than previously possible.

**Crowdsourcing for Data Processing** While the ability to generate sensor data on mobile devices has increased dramatically, our ability to process and filter this data to extract useful, actionable information is still a major challenge. Sensor data from a phone presents several data quality challenges such as blurry, dark, or out-of-focus images, high background noise and clutter in audio, GPS error, time-varying sensor orientation, sensor calibration issues, improper sensor placement, missing samples, and others. In addition to these intrinsic quality issues with data from an individual device, data collection from the masses presents challenges due to the need to filter redundant submissions and spam, and to verify the authenticity of the data. Thus, addressing data quality looms as one of the biggest challenges in handling the torrent of sensor data from mobile devices.

We argue that one of the limitations of existing mobile data processing systems is that they are focused solely on the "computation" aspect *i.e.* more sophisticated data processing techniques, and the use of large computing resources in the cloud. In contrast to automated processing, humans are surprisingly good at filtering sensor data and identifying the most relevant information. Crowdsourced data processing is particularly effective when humans are better than existing automated computer algorithms, for example labeling images, transcribing audio, and others. It is precisely this ability that we seek to tap into in-order to design a human-in-the-loop mobile sensor data processing system.

While cloud computing and crowdsourcing have been used largely separate of one another, we believe that they are more powerful in conjunction than isolation. In particular, we believe that new mobile services will involve tight integration of clouds and crowds in a feedback-driven manner — where clouds use state-of-art algorithms to process sensor data, but use crowds when their confidence in the result is low — and the crowds provide feedback to the clouds on the quality of results to enable continuous improvement in algorithm parameters. We envison an integrated information architecture that combines clouds and crowds to enable in-depth, large-scale, and cost-effective information gathering, and more accurate techniques for information extraction from data.

Our vision is to develop and implement a comprehensive data quality assessment, filtering, cleaning, aggregation, and validation framework that combines sophisticated computational data processing in conjunction with human computation, and that can be used on a wide range of mobile services such as participatory sensing, mobile multimedia search and mobile healthcare.

*Author Bios and Expertise:* The participants are faculty members at the University of Massachusetts Amherst, and have extensive experience with wireless sensing, mobile computing, image search and processing, and applications. Ganesan is a program co-chair for ACM SenSys 2010, and Corner is a program co-chair for ACM MobiSys 2011, the two top conferences in sensor and mobile systems.