CrowdSignals: A Call to Crowdfund the Community's Largest Mobile Dataset

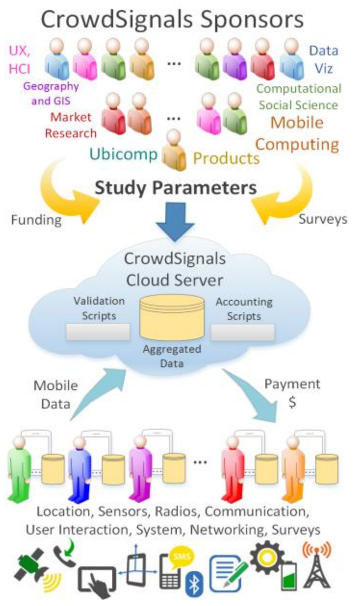


Figure 1: Diverse Sponsors of the CrowdSignals Campaign build funding and surveys in exchange for rich mobile data from 100s of smartphones.

Evan Welbourne
CrowdSignals Campaign
San Francisco, CA 94117 USA
evan@crowdsignalscampaign.com

Emmanuel Munguia Tapia
CrowdSignals Campaign
San Francisco, CA 94117 USA
emmanuel@crowdsignalscampaign.com

UbiComp '14, September 13 - 17 2014, Seattle, WA, USA
Copyright 2014 ACM 978-1-4503-3047-3/14/09…$15.00.
http://dx.doi.org/10.1145/2638728.2641309

Abstract

Researchers from diverse backgrounds critically depend on mobile datasets. From training and testing activity recognition models, to verifying hypotheses in social science, mobile data is indispensable. Unfortunately, mobile data collection requires significant time and budget for infrastructure as well as subject recruiting, screening, training, legal agreements, equipment, and compensation. We estimate up to 70% of the resources in a study may be spent on data collection. Moreover, this massive investment can combine with institutional, legal, and political issues to create a disincentive to sharing with the community. In this paper, we propose and justify a crowdfunded and crowdsourced methodology for longitudinal mobile data collection that cuts researcher costs by orders of magnitude, removes barriers to data sharing, and boosts data value for all stakeholders. We also present CrowdSignals, a first instantiation which will generate the largest labeled mobile dataset available to the community.

Introduction

With a global penetration rate over 96% [5], a majority of people own mobile phones. In developed urban areas where the rate is 100%, everyone owns a smartphone and other devices - if not, they will before 2020 [8]. These devices capture our shared thoughts and feelings along with rich streams of sensor, social, and system data. Data on our activities, environment, mobility patterns, preferences, interactions, social life, and software usage are all latent in this *mobile data*. As such, mobile data can be used to recognize, track, and react to practically any phenomena in our proximity, be it physical, psychological, or sociological.

Consequently, mobile data is integral to research and in a variety of fields. The most valuable data are collected longitudinally from many subjects in naturalistic settings because it produces more robust results. The value of a data set increases with each logged data type, and it increases dramatically more when subjects *label* their data with their current activity, feeling, opinion, place, social situation, or surroundings. The labels can later be used by machine learning algorithms to find interesting correlations or to classify events

Unfortunately, data is scarce and mobile data collection is expensive, time consuming, and difficult. In this paper, we propose and justify a crowdfunded and crowdsourced methodology that cuts costs by orders of magnitude, facilitates sharing, and increases value for all stakeholders. We also introduce the CrowdSignals campaign (see Figure 1), the first instantiation of this approach which will create the largest labeled mobile dataset available to the community.

Mobile Data Collections

Several longitudinal mobile data collection campaigns since 2000 are summarized in Table 1 with respect to scale, date, duration, and estimated cost. Each accelerated research in important ways and represents a different combination of methodology, cost, and experimental control. In this section we present and contrast several key aspects of these campaigns.

Campaign	Subjects	Date	Length (mo.)	Cost (US $)
MIT Reality Mining [2]	100 Students	2004	9	$110K+
Dartmouth CenceMe [10]	9000+ People	2008	N/A	$75K+
Nokia LDCC [11]	< 200 students	2009	24	$1.2M+
MIT Social fMRI [1]	< 200 students	2011	9	$150K+
UB PhoneLab [14]	200 students	2012	12-48	$1.6M+
SMU LiveLabs [15]	30K people	2013	12+	$20M+
Samsung CS [17]	63 people	2013	3	$15K

Table 1. Mobile data collection campaigns since 2000.

Methodologies

Three key methodologies are: local administration (LA), app store-based (AS), and crowdsourced (CS). Most common is LA in which researchers painstakingly recruit, on-board, and manage subjects in person, compensating them with smartphones and a mobile data plan. The MIT, Nokia, UB, and SMU campaigns are examples of this approach. CenceMe used AS in which an app is distributed to 1000s of users via an app store. In the CS approach [17], remote subjects are rapidly recruited, managed, and paid with a crowdsourcing service, but they install a data collection app on their own phone. Our methodology builds on this with crowdfunding.

Contributions

Low-Cost: crowdfunding amortizes expenses over a big community, cutting costs by orders of magnitude.

Sharing: crowdfunding means that all interested parties are engaged from the start, so data sharing is a core component of the campaign from its inception.

Scale: the crowd generates a larger pool of funds than the individual, enabling larger scale data collection.

Customization: sponsors specify surveys with which subjects label the data, customizing and adding significantly more value.

Ethical: by following best practices for informed consent, user agreements, privacy, security, and data sharing we match the standards held by most institutional IRBs.

In addition, we present and explain the CrowdSignals Campaign as a concrete example of this approach.

Collected Data

Location and Radios: Bluetooth, GPS, GSM, WLAN (30-60 sec, every 10 min)

Sensors: Accelerometer, Ambient Temperature, Gravity, Gyroscope, Light, Magnetic Field, Microphone, Orientation, Pressure, Proximity, Humidity, Rotation (10 sec, every 5 min)

Social: Calls, Contacts, SMS; sensitive content is hashed (every 5 hours)

System and Networking: Battery, Connections, Network Traffic (every 5 min)

User Interaction: App Installs, App Launch/Close, Browser Logs, Phone Settings, Configuration (event-triggered)

Subject Feedback:
Lockscreen Questions
(every screen unlock)
ESM Questionnaires
(1-2 times per day)
Entry/Exit Surveys
(beginning and end of study)

Figure 2: List of collected data types along with the sampling window size (if applicable) and sampling intervals.

	Software	**Hardware**	**Admin**	**Comp**
Local Admin	$2000 x 4 months x # developers	$300-$600 x # subjects	$2000 x # months x # assistants	$50-$100 x # months x # subjects
App Store	$2000 x 8 months x # developers	N/A	$N/A	N/A
Crowd-source	$2000 x 4 months x # developers	N/A	$500 x # months x # assistants	$20-$50 x # months x # subjects

Table 2. Estimated US$ cost for methodologies, including: Software Development (grad students), Hardware (phones), Compensation (cash, data plan), and Admin (grad students).

Cost Breakdown

Campaigns incur costs in several ways (see Table 2). First is developer salary for the app and server. This is more for the high-fidelity AS app. LA also incurs the cost of subject smartphones. A third cost is admin staff; AS avoids this and CS minimizes it. The last cost is payment, which includes data plan for LA studies and is eliminated in AS. Estimates in Table 2 assume grad students are developers and admins. Since many data analyses take 3-4 weeks, mobile data collection may consume upwards of 70% of a research budget.

Experimental Control

Methodologies differ in terms of control. AS releases an app into the wild to see what data comes back. In contrast, LA emphasizes face-to-face recruiting and admin with complex control (e.g., interventions). CS studies such as CrowdSignals lay between: admins interact with remote subjects, but surveys and simple protocol changes can be used to enhance control.

Crowdfunding

In crowdfunding, a group of *sponsors* pay to *fund* a project in exchange for some *reward*. Here we discuss the anatomy of a crowdfunding campaign for mobile data collection using CrowdSignals as an example.

Sponsors

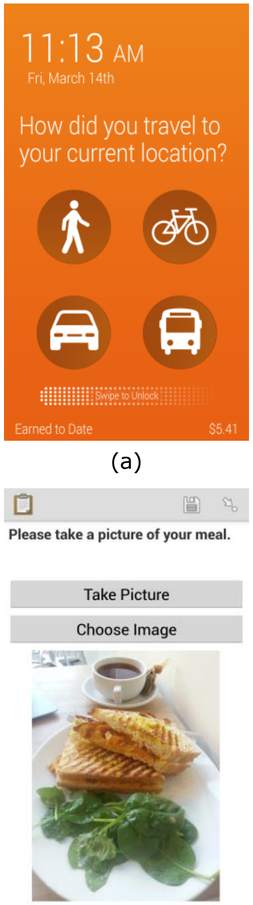
Sponsors come from any field that uses large mobile datasets (e.g., geography, health, sociology, ubicomp). For example, a professor may sponsor because she needs a large sensor data set from diverse subjects, labeled with activities like driving a car or riding a bus. The sponsors split the total cost and because no single institution bears the total expense, there are fewer financial and political barriers to sharing the data.

Funding

Crowdfunding campaigns set a funding goal that will allow the organizers to execute the project. For mobile data collection this should cover all costs, scaled to the size and duration of the study. The early CrowdSignals goal is $50K for 250 subjects over 4 months, with $30K devoted to compensation and for development and admin. Based on early success, CrowdSignals may set *stretch goals* that significantly increase the number of subjects and the duration of the study.

Rewards

Crowdfunding offers rewards based on sponsorship. We reward sponsors with *data*. However, richness of data and allowable usages vary by sponsorship level. *Tentatively*, CrowdSignals sponsors receive anonymous logs (e.g., apps, calls, SMS), mid-level sponsors also receive sensor data and surveys, higher levels receive all data including more sensitive logs (e.g., location) and may specify custom surveys (see next Section).



How did you travel to your current location?

(a)

Please take a picture of your meal.

(b)

Figure 2: (a) Lock screen feedback for frequent, lightweight labels. (b) Open Data Kit Survey for configurable, more in-depth ESM questionnaires.

System Architecture

We use an extended version of the architecture presented by Welbourne et al. [17]. This includes a crowdsourcing service, an Android data collection app, and a cloud server. The section describes the CrowdSignals architecture as an example.

Crowdsourcing

Recruiting, management, and payment are conducted on crowdsourcing platforms Elance [3] and ODesk [12]. We will also recruit external subjects (e.g., through online or physical ads) into the platform where they can be more easily managed. Subjects will interact with a team of trained study administrators during the study.

Data Collection App

The Android app has a background service, configurable surveys, and basic controls. The service launches on boot, collecting the data in Figure 2. All sensitive data (e.g., SMS content) is hashed to protect subjects. The service securely uploads encrypted, compressed data when the subject's phone connects to WLAN.

Sponsors may solicit ground truth from subjects with a survey framework that offers lockscreen surveys and experience sampling method (ESM) [7] surveys (see Figure 2). The lockscreen solicits frequent, lightweight feedback with a multiple choice question (e.g., "How do you feel?") every time subjects unlock their phone - about 19 times per day on average [16]. ESM questionnaires use the Open Data Kit (ODK) [13] for configurable surveys and participatory sensing (e.g., audio, video, barcodes). Top sponsors may specify custom ESM questionnaires using ODK's JSON survey specifications; lockscreen surveys are specified as a combination of text and image files.

Basic controls allow subjects to start and stop all data collection. Subjects may also review the amount of data and survey responses they have uploaded. Finally, subjects may explicitly upload their data when connected to WLAN using an "Upload Now" button.

Cloud Server

The cloud server includes an HTTP server that accepts secure uploads from the clients, and batch scripts that post-process received data. The scripts ensure the fidelity of the data and track how much each subject has uploaded. This allows us to pro-actively contact subjects when problems arise and to pro-rate payment based on how much data each subject has contributed.

Study Protocol

Figure 3 shows the end-to-end flow of a crowdfunded, crowdsourced data collection campaign which we describe in more detail below.

Funding and Requirements Gathering

Funding is organized using a crowdfunding platform [6,9]. During the funding period, input from sponsors regarding surveys and target subject demographics must be collected and reviewed as well.

Recruiting and On-Boarding

Study administrators begin recruiting and on-boarding by inviting candidates to apply on the crowdsourcing platform – this may include candidates that are not already on the platform. Admins then screen applicants and walk them through informed consent, after which subjects receive instructions on how to install and use the app. Finally, subjects complete entry surveys collecting information on demographics, personality, preferences, or other information useful to sponsors.

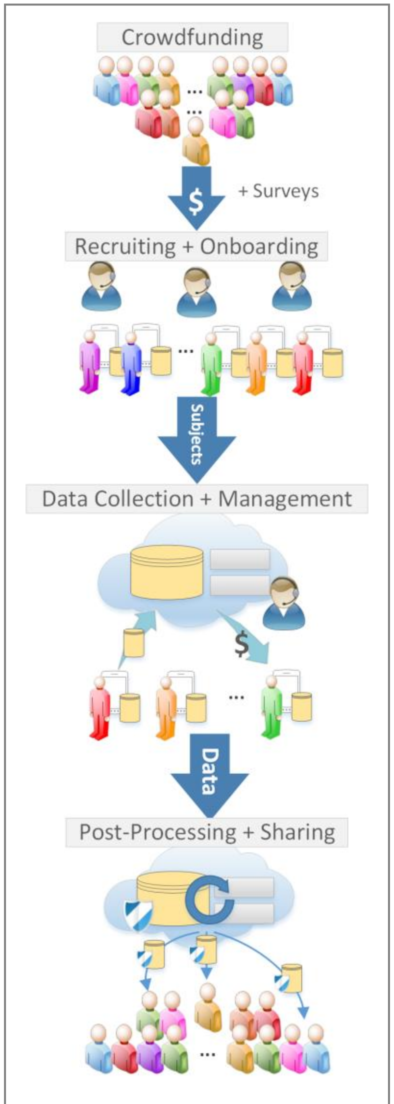


Figure 3: Study protocol

Data Collection and Management

Once subjects are uploading, admins start monitoring collected data with server scripts and solving any problems that arise (e.g., bugs, missing subject data). During this phase subjects may opt-out or may leave the study because they can no longer participate for technical reasons (e.g., switched to iPhone). Subjects may also forfeit their pay to delete all data at any time. Otherwise subjects are paid at the end of the study according to the amount of data they have contributed.

Post-Processing and Sharing

Finally, data is post-processed and shared to sponsors. Watermarks are applied to track any leaks. Sponsors sign a data sharing agreement on ethical use and best practices for storage and access. This includes clauses such as: no contacting subjects directly or reverse engineering anonymized data, and no onward transfer. Finally, the dataset is posted online and sponsors are notified so that they can download their copies.

Conclusion

Mobile data is central to research, but it is expensive and there are many barriers to use. We presented a crowdfunded, crowdsourced methodology that cuts cost by orders of magnitude and reduces barriers to sharing. We also introduced an extensible survey framework for sponsor customization. As an example, we presented CrowdSignals, which if funded, will produce the largest labeled, longitudinal mobile dataset. We invite participation through both critique and sponsorship.

References

[1] Aharony, N. et al. Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing 7 643-659 (2011)

[2] Eagle, N. and Pentland, A. Reality mining: sensing complex social systems. PUC, Vol. 10 Issue 4, pp. 255-268 (2006)

[3] Elance. http://www.elance.com

[4] Funf | Open Sensing Framework http://funf.org/

[5] ICT. The World in 2014, ICT Facts and Figures. ICT http://www.itu.int/en/ITU-D/Statistics/Pages/facts/

[6] IndieGogo. https://www.indiegogo.com/

[7] Intille, S. S., et al., Eliciting user preferences using image-based experience sampling and reflection, ACM CHI '02, pp. 738-739 (2002).

[8] Khan, S. and Marzec, E., Tech Trends 2014: Wearables. Deloitte University Press, http://www.dupress.com/articles/2014-tech-trends-wearables

[9] Kickstarter. https://www.kickstarter.com/

[10] Miluzzo, E., et al. Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. ACM *SenSys (*2008).

[11] Nokia Lausanne Data Collection Campaign, https://research.nokia.com/page/11367

[12] ODesk. http://www.odesk.com

[13] Open Data Kit. http://www.opendatakit.org

[14] Phone Lab: A Programmable Smartphone Testbed http://www.phone-lab.org/

[15] LiveLabs, http://centres.smu.edu.sg/livelabs/

[16] Vaish, R., et al. Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time. CHI'14 (2014)

[17] Welbourne, E., et al. Crowdsourced Mobile Data Collection: Lessons Learned from a New Study Methodology. HotMobile'14 (2014)