ALGOSNAP

CrowdSignals.io Pilot Dataset Documentation

Smartphone, Smartwatch, and Survey data

Reference Document

11/01/2016

	EMAIL	WEB
111 S Jackson St, Seattle, WA 98104	info@AlgoSnap.com	www.AlgoSnap.com

Overview

Motivation

Smartphones and smartwatches are witness to detailed information about practically every aspect of our lives as individuals and communities - from our sleep, to our health and fitness activities, to our social life, media consumption, and mobility patterns. Mobile data are increasingly one of the largest drivers of innovation in areas as diverse as computing, Internet of Things (IoT), geography, journalism, marketing, mHealth, mobile computing, sensing, social science, and urban planning among others. Indeed, personal data are often referred to as "The New Oil" for the 21st century [5] - and several pioneering projects in the last decade (e.g., Nokia Mobile Data Challenge [2]) have demonstrated that precise, ethically collected data from personal devices can supply pertinent, never-before-available information on the specific behaviors, patterns, and trends of individuals, groups, and our society as a whole.

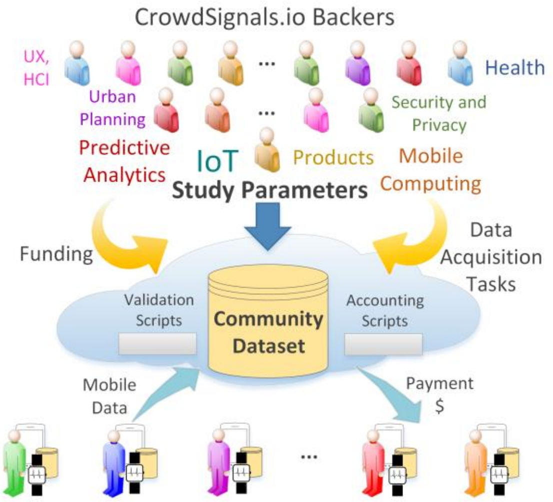

Figure 1 Crowdfunding collection of a shared dataset

Despite the intense focus, there's a critical problem for academia and businesses alike: smart device data collection campaigns are extremely expensive, time consuming, and challenging both technically and legally. In addition, many organizations lack the time, funding, and human resources to collect high-quality labeled data from a diverse population. As a result, data are incredibly scarce - and those who endeavor to collect them end up spending an enormous amount of time and money on data collection infrastructure, legal services, administration, and management.

CrowdSignals.io is an initiative that begins to address the data problem by unifying the research community with a campaign to crowdfund a shared mobile dataset. The initiative collects rich, high-quality smartphone and smartwatch data from a diverse group of participants for 30 days. Crowdfunding with support from hundreds of researchers ensures that the cost of the dataset is orders of magnitude less than an in-house data collection. By collecting over 40 types of sensor, social, system, and user interaction data, CrowdSignals.io generates a large dataset that is widely applicable to research in many fields.

CrowdSignals.io Crowdfunding Campaign

Formative feedback from researchers and scientists were gathered from October to December 2015. Online surveys were sent to mailing lists as well as LinkedIn and Facebook groups for data science, Internet of Things, mobile health, mobile and ubiquitous computing, and sensors. Survey data helped to identify the types of data, participants, and participant feedback that were most valued by prospective Backers. Table 1 presents results from 249 survey responses in the form of top-5 lists for several categories.

In addition to surveys, more than 20 experts from academic and industrial research institutions were consulted for their opinions and advice regarding the CrowdSignals.io campaign. These discussions helped to focus the proposed data collection while maintaining broad appeal to researchers in a variety of areas.

Table 1. Top-5 lists from survey feedback on most valuable types of data, ground truth, and participant demographics.

Data	Ground Truth	Demographics
1) Inertial Sensors (e.g., accel, gyro)	1) Precisely labeled data interval	1) 65+ years old
2) Location (e.g., GPS, GSM)	2) Timestamped survey response	2) "Not just students"
3) User Interaction (e.g., app usage)	3) Demographic surveys	3) Diverse ethnicity
4) Environmental Sensors (e.g., light)	4) Participatory sensing	4) Has a specific medical condition
5) Social (e.g., Calls, Facebook, SMS)	5) Ecological Momentary Assessment	5) Balanced sex

Table 2. Parameters for the data collection as proposed in the crowdfunding campaign.

Data	Inertial and environmental sensor data, anonymized user interaction and social data
Devices	Android Smartphones and Smartwatches
Feedback	Labeled intervals, Timestamps surveys, Demographic Surveys
Participants	30 participants with diverse age, sex, education, and ethnicity
Duration	30 days

The crowdfunding campaign parameters and promotional materials were developed in January 2016. Community feedback on CrowdSignals.io were applied to maximize value of the proposed data collection for the research community under financial, legal, and privacy constraints. As Table 2 presents, the proposed data collection captured the most valuable types of data, feedback, and participant demographics while omitting the more sensitive, expensive, and legally challenging types (i.e., location data, participants with medical conditions).

The principal mechanism in the crowdfunding campaign was to allow individuals and groups to contribute funds and become Backers in exchange for a license to the collected data. The funding goal was set to $15,000 USD, the estimated cost of the proposed data collection given an existing and ready-to-use software and legal framework. Specific contribution levels and perks were designed to make the collected data accessible to individuals and small groups with limited funds while attracting larger contributions from larger institutions. Promotional materials included a website and crowdfunding video that describe and endorse the initiative with basic graphics and comments from numerous experts.

Campaign Results

The campaign was successful, achieving 160% of the funding goal with a total of $24,066 USD raised from 164 Backers. Indiegogo contributions accounted for $20,778 ($18,066 after 7% transaction fees) of funds, and an additional $6000 came directly from institutional sponsors. Table 3 breaks down the Indiegogo contributions by level. Surveys were sent to Backers to learn their self-declared area of expertise in addition to the application area in which they would apply the CrowdSignals.io data. The top areas of expertise were IoT, Data Science, Ubicomp, Sensors, and Networks/Systems. Nearly half the funds came from IoT start-ups that needed data with a specific ground truth label. The next largest block of funding came from individual data scientists and data science institutions that were interested in exploring the dataset. The most common applications to which the data would be applied were Predictive Analytics (e.g., predicting user preferences or behavior), Education, and Health (primarily mHealth).

Table 3. Indiegogo funds raised by contribution level

Level	Backers	Funds Raised	Subtotal
Thanks!	4	$40	$40
Academic	73	$1460	$1500
Academic Profile	15	$450	$1950
Commercial	8	$400	$2350
Commercial Profile	5	$300	$2650
Educational	6	$360	$3010
Academic Sponsors	7	$700	$3710
Commercial Sponsors	4	$1000	$4710
Academic Groups	5	$1000	$5710
Commercial Groups	0	$0	$5710
Guarantee your label!	5	$12,500	$18,210
Specials	17	$850	$19,060
Donations	14	$1718	$20,778

Email and survey feedback were received from Backers and members of the research community during and after the crowdfunding campaign. Several insights were derived from this feedback. First, the purchase approval chain at both academic and commercial institutions can take weeks and often does not permit or reimburse crowdfunding contributions. The campaign was extended to 60 days to allow approval chains to be processed. Contributions from two institutions were accepted directly when payment through Indiegogo was not possible. Contributions from several larger commercial institutions were canceled due to the time and expense of iterative legal revisions to the data license - both for the institutions and for AlgoSnap.

Backers at some academic institutions needed internal review board (IRB) or ethics committee approval before contributing. For this purpose, an Ethics and IRB reference document was prepared to describe participant recruiting, informed consent, and compensation procedures in addition to techniques applied to ensure security, privacy, and anonymity of participants.

Survey responses from 131 community members revealed the importance of several criterion to the decision to become a Backer or not. Responses were provided as a rating on a 1-5 Likert scale with 1 being “Not Important” and 5 being “Critically Important”. The top five most important criterion with weighted average score were: “Availability of a certain type of ground truth” (4.18), “Availability of a data from certain sensors or devices” (3.76), “Availability of sample datasets” (3.41), “Expertise and reputation of campaign organizers” (3.38), and “Cost of a contribution to receive data” (3.0). Community members that did not back the campaign most often provided one of several reasons: they did not know about the campaign in time to support it, a US-based data collection campaign was not interesting to them, or they were not able to obtain reimbursement from their institution for a crowdfunding contribution.

CrowdSignals Platform

The CrowdSignals platform comprises mobile, wearable, and IoT edge software with accompanying cloud services and legal framework for privacy-sensitive capture of precisely labeled sensor, social, system, and UX data. Researchers have built many excellent frameworks for mobile data capture [1, 8]. AlgoSnap used its proprietary CrowdSignals platform for features that rapidly configure and deploy simultaneous data capture objectives in conjunction with a robust legal framework; this section describes those features.

Configurable Data Capture

The CrowdSignals platform includes an app that captures data from more than 40 *Sources* on smartphones and smartwatches as shown in Table 7. Each Source is configurable with respect to the parameters of the device in addition to sampling rates, duty cycles, and smart hub parameters for high frequency sensors.

Source Data Format

Source data are captured in streams using Apache Avro for compact, efficient data storage and transmission. Streams from periodic Sources (e.g., accelerometer, gyroscope) are structured as Windows of precise duration that are subdivided into Panes and may be separated by an arbitrary sleep period for duty cycling as shown in Figure 2. Aperiodic events (e.g., battery level changes, screen state) are captured as Avro objects with a single event timestamp. All timestamps are recorded in nanoseconds.

Table 6. Supported smartphone and smartwatch data Sources

Android Phones	
Mobility and Radios	CDMA, GSM, LTE, WCDMA, WLAN, Place Visits
Sensors	Accelerometer, Linear Acceleration, Gravity, Gyroscope, Uncalibrated Gyro, Humidity, Light, Magnetic Field, Pressure, Proximity, Rotational Vectors, Significant Motion, Steps, Temperature
System and Network	Battery, Connection, Connectivity, Network Traffic
User Interaction	Phone State, Screen, App Usage
Social	Anonymized call and SMS logs
Android Wear Devices	
Sensors	Accelerometer, Gyroscope, Magnetic Field
Microsoft Band 2	
Sensors	Accelerometer, Altimeter, Ambient Light, Barometer, Calories, Contact, Distance, Galvanize Skin Response, Gyroscope, Heart Rate, Pedometer, RR Interval, Skin Temperature, UV Exposure

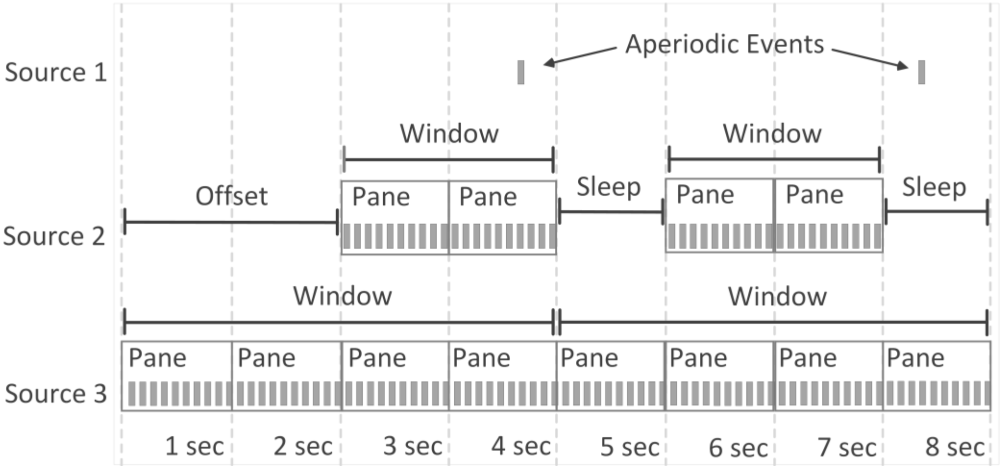

Figure 2 Periodic data Sources may be configured in Windows that divide into Panes and may be separated by a sleep period. Aperiodic events such as screen events or passive location updates are received, encoded, and stored within individual Avro objects.

Source Groups

The platform configures data capture in *Source Groups*, sets of one or more Sources that start and stop together. Sources within a Source Group may be configured independently but are started together and aligned against a global start timestamp - although any Source may be configured to start with an arbitrary offset. Varying offsets allows the user to create staggered or alternating patterns among the Sources within or across Source Groups. Multiple Source Groups may run simultaneously and include the same Sources - even if the shared Sources have different parameters in each group (e.g., different sample rates or duty cycle for the same sensor). All data capture is configured with Source Groups that are specified at run time through JSON configuration files or remotely by a server.

Configurable Data Acquisition Tasks

The CrowdSignals platform supports a variety of Data Acquisition Tasks (DATs) for smartphones and smart watches. Three types of smartphone-based DAT were used to capture ground truth from data collection participants: labeled interval tasks, lockscreen survey tasks, and ecological momentary assessment tasks. The CrowdSignals platform also supports smartwatch-base DATs which allow initiation and control of data acquisition via smartwatches. Each type of task can be rapidly configured with JSON and created from a file or remotely from a server.

Labeled Interval Tasks

Interval labels were the most requested ground truth in survey feedback from Backers and the wider community. An interval label consists of participant-initiated start and end timestamps along with metadata about the subject phenomenon. For example, an interval label for “shaving” could include a start and end time as well as metadata on devices, where each device was carried or worn, and the type of razor used (e.g., electric). Participants explicitly create interval labels by selecting a phenomenon to label, completing a brief metadata survey and clicking a “start” button. The participant then captures the phenomena accordingly, after which they click an “end” button (or a “cancel” button if they made a mistake). A labeled interval task specification may include text or video instructions regarding the phenomenon to be captured. Each interval label task has an associated Sensor Group that is started when the participant clicks “start” and stopped when the participant clicks “end” (e.g., the “Shaving Source Group” which captures continuous accelerometer data from the watch). Participants can also modify some survey metadata (e.g., a change in placement of the phone) during an interval label capture if they needed to.

Lockscreen Survey Tasks

Lockscreen surveys are an effective technique for gathering many points of light weight feedback from participants throughout the day [9]. This task presents participants with a single-screen survey when they unlock their phone. Lockscreen surveys must take at most 5 seconds to complete, and on submission a single timestamp is recorded along with the participant response. For example, a lockscreen survey may ask a participant "What kind of place are you currently in?" and present a drop-down list of place categories. Participants may also skip lockscreen surveys any time they are presented by clicking a "skip" button.

Ecological Momentary Assessment Tasks

Ecological Momentary Assessment (EMA) is a survey technique from behavioral medicine research whereby a study participant repeatedly reports on a particular phenomenon (e.g., providing a label) close in time to experience and in the participant's natural environment. The CrowdSignals Platform supports EMA-style labels that are triggered at random intervals throughout the day or by a particular event in the participant's context (e.g., a geo-fence around their home or work). For example, an EMA task many ask a participant to describe their mood using the 2D circumplex model [4] at random times throughout the day.

Mixed Type Tasks

Interval, Lockscreen, and EMA labeling may be combined and mixed to achieve the desired density, precision, and cost of a set of ground truth labels. For example, a Lockscreen or EMA survey may present an Interval Labeling task.

Cloud Storage and Processing

The CrowdSignals app periodically compresses and uploads the collected Source and DAT data to the cloud. By default, uploads occur only when the phone is charging and connected to Wi-Fi, but the software may be configured for uploads over LTE (e.g., if a participant has an unlimited data plan). In the cloud, data are periodically validated and analyzed with an Apache Beam pipeline to assess the quality and quantity of Source and DAT data provided by each participant.

Consent and Privacy

In addition to the rigorous person-to-person informed consent process described below, the app also walks participants through an e-consent process. In addition to granting permission to use each Source during install (e.g., Body Sensors, Heart Rate Sensor), participants read and approve a plain English request to use Sources.

Participants may also turn off data collection and upload at any time using a simple switch in the app's administration interface. The participant legal agreement also allows participants to opt out, uninstall the app, or request that AlgoSnap staff destroy all their data at any time without penalty. By default, Source data that include unique identifiers or other potentially sensitive information (e.g., WLAN MAC addresses, content and contacts for Calls and SMS) are one-way hashed on device and then replaced with generic but unique identifiers (e.g., "access point 754", "call to contact 23 at 2:03pm").

Please see the CrowdSignals.io Ethics and IRB reference document for additional details.

Pilot Data Collection

The data collection occurred from late August to November 2016. While the collection followed parameters set forth in Table 2, additional details were decided using Backer input and available funds. This section provides those additional details along with preliminary results from the data collection.

Recruiting Participants

Participants were recruited and filtered in a 3-step process that began with online forums, mailing lists, and crowdsourcing sites. Following brief email correspondence, prospective participants were contacted by phone to discuss the collected data, requirements for participation, available compensation, and how the collected data would be used along with the ensuing risks and benefits of participating. Those interested were invited to meet in person for a 1-2 hour information and training session with an AlgoSnap staff member. During the training session, prospective participants had the opportunity to ask questions and, if they consented to participate, signed the participation agreement. Participants then paired a loaned smartwatch with their phone and installed the CrowdSignals app before working with AlgoSnap staff in a 1.5 hour training session. During the training session, participants practiced a set of representative DATs under expert supervision. Participants that collected good quality Source and DAT data were invited to keep the loaned watch and continue collecting data for 4-6 weeks.

The recruiting process accepted 40 participants for on-site training and selected 25 for the long term data collection; 5 additional participants were recruited and trained by participating friends or family members after informed consent through a phone call with AlgoSnap staff. Among the 30 participating in long term data collection were 18 males and 12 females of varying age, height, weight, education, and ethnicity. The included spreadsheet summarizes basic demographic information for these 30 participants.

Smartphones and Smartwatches

The Crowdsignals.io campaign and software platform are currently focused on the Android ecosystem. As such, every participant owned an Android smartphone. However, since the market penetration of Android smartwatches is relatively low, participants were loaned a smartwatch as described in the previous Section. The attached spreadsheet describes all smartphones used - the Samsung Galaxy Line and LG G4 were most common. Concerning distribution of smartwatches, 12 participants used Microsoft Band 2, 9 participants used Asus ZenWatch 1, 1 participant used Asus ZenWatch 2, 1 participant used the Motorola Moto 360, and 5 participants used the Sony 3 watch. As discussed below, participants were incentivized to use smartwatches during the data collection period but were not required to do so and some did not wear their watch at all.

Source Group Configuration

One Source Group called “background sensors” (see Table 7) was configured to run periodically in the background even when participants were not actively performing a DAT. This Source Group included all smartphone and watch sensors listed in Table 6, sampled in 20-second Windows with 1-second Panes and having a 40-second sleep interval between Windows. This is frequent for background sensing but it maximizes the density of information captured, sensor hub technology lowers power consumption considerably when available, and most participants charged their phone multiple times per day by habit. Participants could also manually shut off background sensors anytime and subsets of Sources were omitted if not available on a particular device. In addition, Sources may stop streaming periodically on devices with low resource availability. Other Sources (i.e., mobility and radios, system and networking, user interaction, social) were sampled at much lower rates.

Table 7. Background sensors Source Group which ran continuously during the data collection period so long as participants did not disable it.

Background Sensors Source Group						
Device	Source	Frequency	Window	Pane	Sleep	Offset
Android Smartphone	Accelerometer	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Linear Acceleration	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Gravity	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Gyroscope	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Uncalibrated Gyroscope	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Light	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Magnetic Field	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Pressure	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Rotational Vectors	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Temperature	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Proximity	--	20 sec	1 sec	40 sec	0 sec
	Significant Motion	--	20 sec	1 sec	40 sec	0 sec
	Steps	--	20 sec	1 sec	40 sec	0 sec
	Battery	Always listening for events				
	Connection Strength	Always listening for events				
	Connectivity	Always listening for events				
	Network Traffic	3 scans	--	--	5 min	0 sec
	Cellular radio	Single scan	--	--	1-5 min	0 sec
	WLAN				5 min	0 sec
	Phone State	Always listening for events				
	Screen	Always listening for events				
	App Usage	Single scan	--	--	5 min	0 sec
Microsoft Band 2	Accelerometer	~62 Hz (16ms sleep)	20 sec	1 sec	40 sec	0 sec
	Altimeter	Device default	20 sec	1 sec	40 sec	0 sec
	Ambient Light	Device default	20 sec	1 sec	40 sec	0 sec
	Barometer	Device default	20 sec	1 sec	40 sec	0 sec
	Calories	Device default	20 sec	1 sec	40 sec	0 sec
	Contact	Device default	20 sec	1 sec	40 sec	0 sec
	Distance	Device default	20 sec	1 sec	40 sec	0 sec
	Galvanic Skin Response	Device default	20 sec	1 sec	40 sec	0 sec
	Gyroscope	~62 Hz (16ms sleep)	20 sec	1 sec	40 sec	0 sec
	Heart Rate	Device default	20 sec	1 sec	40 sec	0 sec
	Pedometer	Device default	20 sec	1 sec	40 sec	0 sec
	RR Interval	Device default	20 sec	1 sec	40 sec	0 sec
	Skin Temperature	Device default	20 sec	1 sec	40 sec	0 sec
	UV Exposure	Device default	20 sec	1 sec	40 sec	0 sec
Android Wear Device	Accelerometer	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec
	Gyroscope	SENSOR_DELAY_FASTEST	20 sec	1 sec	40 sec	0 sec

Each interval label DAT also had a corresponding Source Group that was triggered to run (simultaneously with the background sensors Source Group) whenever the participants clicked “start”. As shown in Table 8, all Source Groups corresponding to interval labels were configured to run all smartphone and smartwatch sensors in Table 6 at the max sampling rate in 4-second Windows with 1-second Panes and having no sleep interval between Windows. It’s notable that the high data rate from diverse, simultaneously running sensors can degrade the user experience and even the sensor data quality on some devices. However, this configuration was selected because the density of information delivered by a diverse mix of high frequency sensor data was prioritized over optimal sensor data quality by Backers in survey and email feedback.

Table 8. DAT Source Groups which ran whenever a participant began a Labeled Interval DAT and continued until that DAT ended.

DAT Source Groups						
Device	Source	Frequency	Window	Pane	Sleep	Offset
Android Smartphone	Accelerometer	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Linear Acceleration	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Gravity	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Gyroscope	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Uncalibrated Gyroscope	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Light	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Magnetic Field	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Pressure	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Rotational Vectors	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Temperature	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Proximity	--	4 sec	1 sec	0 sec	0 sec
	Battery	Always listening for events				
	Connection Strength	Always listening for events				
	Connectivity	Always listening for events				
	Phone State	Always listening for events				
	Screen	Always listening for events				
Microsoft Band 2	Accelerometer	~62 Hz (16ms sleep)	4 sec	1 sec	0 sec	0 sec
	Altimeter	Device default	4 sec	1 sec	0 sec	0 sec
	Ambient Light	Device default	4 sec	1 sec	0 sec	0 sec
	Barometer	Device default	4 sec	1 sec	0 sec	0 sec
	Calories	Device default	4 sec	1 sec	0 sec	0 sec
	Contact	Device default	4 sec	1 sec	0 sec	0 sec
	Distance	Device default	4 sec	1 sec	0 sec	0 sec
	Galvanic Skin Response	Device default	4 sec	1 sec	0 sec	0 sec
	Gyroscope	~62 Hz (16ms sleep)	4 sec	1 sec	0 sec	0 sec
	Heart Rate	Device default	4 sec	1 sec	0 sec	0 sec
	Pedometer	Device default	4 sec	1 sec	0 sec	0 sec
	RR Interval	Device default	4 sec	1 sec	0 sec	0 sec
	Skin Temperature	Device default	4 sec	1 sec	0 sec	0 sec
	UV Exposure	Device default	4 sec	1 sec	0 sec	0 sec
Android Wear Device	Accelerometer	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec
	Gyroscope	SENSOR_DELAY_FASTEST	4 sec	1 sec	0 sec	0 sec

DAT Configuration

In the month following the crowdfunding campaign, Backers specified additional requests for the type of participant feedback they wanted in the dataset. Backers contributing at the "Guarantee your label!" level were able to choose a particular type of ground truth: 4 chose a labeled interval for a particular activity, 1 chose Lockscreen survey feedback, all requested that their ground truth be kept confidential and not be shared with other Backers – as such it is not disclosed here. Backers at other levels proposed the types of ground truth that would be most interesting to them; proposals were aggregated into DATs that are summarized in Tables 9 and 10. Some Backer-proposed DATs (e.g., driving with GPS, person falling) were not captured due to cost, privacy, or liability constraints

Interval Label DATs

Participants were instructed to collect interval label DATs as they performed these activities in daily life; incentives were offered per label as described below. Among captured Interval Label data were: the Start timestamp of the interval and the corresponding End or Cancel timestamp, metadata on the placement of the phone and watch, and potentially other metadata on the phenomenon being labeled (e.g., "up" or "down" in an elevator). Participants could modify the additional metadata mid-flight to reflect changes in the captured phenomenon. For example, a Participant could start a Walking label with the phone in her back right pant pocket but move it to her left jacket pocket during label capture and reflect that change in the metadata.

Table 9. Labeled Interval DATs performed by data collection participants.

Interval Label	Description	Additional Metadata
Riding bus	Start on entering bus, End just after stepping off bus or during the bus ride if the Participant chose to conserve device resources.	Placement of phone, watch
Riding train	Start on entering train, End just after stepping off train or during the train ride if the Participant chose to conserve device resources.	Placement of phone, watch
Riding light rail	Start on entering light rail, End just after stepping off light rail or during the train ride if the Participant chose to conserve device resources.	Placement of phone, watch
Riding ferry	Start on boarding ferry, End when ferry docks at arrival or during the train ride if the Participant chose to conserve device resources.	Placement of phone, watch
Riding in a car	Start on entering and sitting down in car, End just before stepping out or during the car ride if the Participant chose to conserve device resources.	Placement of phone, watch
Riding a bicycle	Start just before mounting bicycle, End just after stepping off bicycle.	Placement of phone, watch
Riding an elevator	Start on entering elevator, End just after stepping out of elevator	Placement of phone, watch
Riding an escalator	Start just before boarding escalator, End just after stepping off escalator	Placement of phone, watch
Riding a Scooter	Start just before mounting razor scooter, End after stepping off Scooter	Placement of phone, watch
Walking	Start on beginning to walk, End when done walking or during the walk if the Participant chose to conserve device resources.	Placement of phone, watch
Walking on stairs	Start on beginning to walk on stairs, End when done walking on stairs or during the walk if the Participant chose to conserve device resources.	Placement of phone, watch Going up or down
Drinking water	Start just before beginning a session of Drinking (e.g., when sitting down to a meal), Participant then periodically Drinks with the hand wearing the watch, using the smartphone to add a "drink at mouth" timestamp to the metadata every time the drink is at their mouth. End when the Participant is done drinking or if the Participant chose to conserve device resources.	Placement of phone, watch Timestamp every time drink is at the Participant's mouth
Playing video game	Start on beginning to play a console video game, End when game is done or during the game if the Participant chose to conserve device resources.	Placement of phone, watch

Lockscreen DATs

Lockscreen DATs were selected and presented whenever Participants unlocked their phone. The selection was random with bias toward selecting a current place category DAT 30% of the time while other surveys selected with equal chance. Participants could skip Lockscreen DATs anytime or disable them altogether if they wished. Lockscreen DATs had no associated Source Group and instead may be analyzed in conjunction with the Background Sensors Source Group. Lockscreen DATs took at most 5 seconds to complete and are presented in Table 10. In addition to the Participant's response, the Lockscreen DAT recorded additional information such as whether the Participant interacted with each UI component (e.g., a dropdown box) or not. Lockscreen DATs were also configured to assist Participants' in rapid response by placing the most recently used answers (e.g., MRU drop-down list item) as the default.

Table 10. Lockscreen DATs performed by data collection participants.

Target Information	Description
Current Place	Asks the Participant about the type of place they are currently in as well as the primary mode of transportation they used to arrive at this place: They respond to the place question by selecting an item from a drop-down list: Home, Work, In transit, Restaurant, Café, Shop or Store, Bar or Nightlife, Supermarket/Grocery, School, Library, Museum, Performance Venue, Tourist Attraction, Fitness/Sports Facility, Bank, Government Building, Outdoors / Park, Traffic Stop/Rest Area, Hospital/Clinic, Hotel/Motel/Hostel, Transit Station, Church, Other. They respond to the mode of transit question by selecting an item from a drop-down list: Bicycle, Bus, Car, Ferry, Light Rail, Train, Taxi/Uber, Scooter, or Walking.
Mood and Physical Wellbeing	Asks the Participant to specify their mood and overall physical wellbeing: Mood is specified using the 2D Circumplex Model with two sliders ranging between 0 and 100 with the first slider extremes labeled "Sleepy" and "Wide Awake" and the second labeled "Unpleasant" and "Pleasant". Overall physical wellbeing was specified using a sliders ranging between "Sick" (0) and "Great" (100).
Phone Position	Asks the Participant where their phone was just before they unlocked it? They respond by selecting an item from a drop-down list: In hand, In bag or purse, On table, Jacket right pocket, Jacket left pocket, Right cargo pants pocket, Left cargo pants pocket, Shirt breast pocket, Right clip, Left clip, On charging station, or On other furniture or surface.
Sedentary Activity	Asks the Participant two questions: 1) Are you sitting or lying down? They respond yes or no 2) If so, what are you doing? They label their current activity using a drop-down box as: Working on a computer, In transit, Socializing, Watching Television, Reading, Thinking, Working, Hobby, Playing a Game, Resting, or Other.

Compensation and Incentives

Participants were compensated $20/hour for their time during the training and consent process. During the data collection period, participants were incentivized to collect data with additional micropayments based on the amount of data and ground truth they contributed. Micropayments were awarded as follows. Participants were paid $0.75 per day and $0.25 per day, pro-rated by the fraction of time during which the background sensors Source Group was running on their phone and smartwatch respectively. Labeled interval DATs were paid $0.20 per instance, up to $1 per day and Lockscreen surveys were paid at a rate of $0.05 per response, up to $1 per day. DAT results were validated in the cloud before payment to prevent participants from gaming the system. For example, interval labels needed at least a minimum duration and must not have been captured consecutively within a short amount of time (e.g., 3 walking labels within 5 minutes would be paid as one label, not 3). Participants were paid at the conclusion of the data collection period and had the option to keep the smartwatch they were using as a substitute for the cash compensation.

Collected Data

Participants collected over 150GB of data corresponding to over 13,000 hours. As anticipated, high frequency sensors such as accelerometer and gyroscope were responsible for the largest streams of data. Moreover, due to the background sensors Source Group, more data were captured by Participants with watches and smartphones having more sensors. However, some participants were clearly more engaged than others with respect to data capture and therefore collected significantly more data. For example, one participant collected more than 25GB of data with over 500 MB of binary Avro data each day, while another often turned off background sensors, collecting less than 2GB overall and less than 100MB of data on most days. Regarding smartwatches, 10 participants reported that they only wore their watch 3-5 hours per day or while capturing interval labels. These participants took off their watch at home or when it interfered with an activity such as typing. Of the 12 participants with Microsoft Band 2 devices, 4 reported that their device stopped working within the first 3 weeks of use.

Collected DAT Results

Overall, participants submitted more than 1000 interval labels and over 3000 lockscreen survey responses. This accounts for more than 100 hours of precisely labeled data and more than 7000 hours containing at least one point of participant feedback. As noted above, several participants were "power users", collecting 5 or more interval labels per day and more than 30 lockscreen survey responses per day on average. The most common interval labels overall were "Walking" followed by "Riding the Bus"; the least common were "Riding Bicycle" and "Riding Ferry".

Delays and Gaps in Data

As often occurs in pilot studies, the CrowdSignals.io pilot study encountered several unanticipated events that impacted the data. We describe those delays and gaps in the data below along with what was our approach to compensating for them as well as our takeaways for future data collection campaigns.

Control and Comfort Preferences

Participants were incentivized to collect data with their phone and watch, but they were not required to. Indeed, administration and privacy controls in the app allowed participants to turn off some or all data collection at any time for any reason. A number of participants disabled specific Sources or Source Groups, resulting in a smaller dataset for those participants. For example, some participants (e.g., User12, User31, User41) turned off Lockscreen surveys because they found them to be too cumbersome despite the incentive. Other participants turned off sensors or decided not to pair and wear the watch to save battery (e.g., due to sensor operation or Bluetooth activity). Still other participants were equipped with a watch but chose not to wear it for comfort or fashion reasons. For example, User19 had a Moto360 watch but chose not to wear it because she felt it was too large, uncomfortable, and didn't look good.

Google Play Services Incompatibility

To communicate with Android Wear devices the CrowdSignals app leverages Android Wear communication libraries that are present in newer versions of Google Play Services. Phones and watches running older versions of Android and Google Play Services needed to update before communication was enabled. While participants were trained in how to update their version of Android and Google Play Services, not all took the time to do so, resulting in a lack of watch data for several participants. In the future we will require that participant devices have a minimum version not only for Android but for Google Play Services as well.

Labeling Errors

The event labeling interface on the CrowdSignals app allowed participants to begin an interval label (e.g., "Riding a Bus") and then navigate away from the app before returning to end the interval. About a third of participants reported at least one instance of starting an interval and then forgetting to end it until 1-4 hours later. In such cases, users were trained to "cancel" the label but we received reports of users accidentally ending the label instead of cancelling it. In every reported case, the participant was performing a longer term activity (e.g., "Riding a bus", "Riding in a car").

Participants also reported interruptions to long running interval labels. Interruptions would often consist of incoming calls, SMS messages, or app notifications. For example, the Participant coded as User31 often took ~1 hour bus rides with her phone in her pocket and metadata indicating as such, however, at times her phone would ring or she would take her phone out of her pocket to check an app, thus violating the declared metadata before returning the phone to her pocket. In such cases, the accompanying data on call and SMS logs as well as screen and phone state are a good indicator for interruptions during long-running activities.

Another common type of labeling error was when participants clicked "submit" on a Lockscreen survey instead of canceling when they wanted to skip it. These errors are somewhat more difficult to trace. One piece of information included with the survey data is the wasInteracted field which indicates whether or not the participant actually touched or altered each field in a survey (e.g., mood slider, physical wellbeing slider). This is an imperfect indication however because Lockscreen surveys cache and present the most recently used feedback by default – so a participant may be correctly answering the survey if they submit without changing the feedback when the cached feedback still accurately reflects their true response.

Limited Device Capabilities

There is significant variation in the Android ecosystem with respect to sensor and system features of smartphones and watches. While we tried to recruit participants with higher-end phones, not all participants had devices with sufficient CPU, memory, sensors, and storage to produce good quality data. Moreover, some participants had phones that were already saturated with apps that continually consumed most of the available resources. In such cases (e.g., User9), the CrowdSignals app collected as much data as possible from the available sensors and data sources but the overall amount and quality of data was lower.

Limited Upload Bandwidth

Perhaps the most crucial finding over the course of the pilot was that sporadic and unexpectedly infrequent WLAN access led to gaps in collected data. While all participants had WLAN access at home and/or work on a daily basis, many did not connect their phones to WLAN every day during the study. In some cases this was an accident of habit, in others a life event (e.g., going on holiday/vacation, permanently moving from one apartment to another as User6 and User12 did) led to a period without WLAN access, and in other cases the participant's home WLAN connection was unstable or much lower bandwidth than expected. The net result was that the CrowdSignals app continued to store data locally on the participant's phone until it reached its maximum storage limit – data were lost until that Participant had enough sustained WLAN access to upload the backlog of data.

While the limited bandwidth problem effected several participant logs significantly by reducing the total amount of data received (e.g., User1, User6, User8, User9, User10, User12), we were able to curb the problem for other Participants in several ways. First, many participants had unlimited mobile data plans – so we released a patch that allowed these participants to upload data over WLAN *or* their mobile data connection. We also recommended a protocol for participants to follow after a prolonged period without WLAN access: stop all sensors, connect to WLAN, and let the phone upload overnight. In the future it seems reasonable to implement a context-aware data logging policy which limited sampling rates and lowers duty cycles when WLAN has not been available or when local storage use reaches a certain percent of capacity.

To compensate for limited data from some participants, we increased the duration of the data collection period for other participants with better quality data by 2 weeks. As such, data from a majority of participants extends over 6 weeks rather than 4 weeks.

Lost, Damaged, or Reset Phones

Participant coded as User 42 lost her phone within the first few days of her participation and then restarted several weeks later when she obtained a new phone – this amounts to a significant gap in data and a late completion of the study. The Participant coded as User 34 lost his phone on a bus, continued the study with a replacement phone until he recovered his old phone in lost+found a week later. For the period during which his original Xiaomi phone was missing, all data recorded was using a different phone – his watch, an Asus Zenwatch 1, remained the same.

Microsoft Band 2 Battery Life

While the Microsoft Band 2 device produces a variety of good quality sensor data, its battery life was often not enough to last an entire day. Moreover, at the request of a "Choose your own label" level Backer the Microsoft Band 2 device was configured to sample data continuously when connected to the phone – in this configuration the band lost charge within about 8 hours. After several weeks in this mode, we reverted to the duty-cycled Microsoft Band 2 data collection which extended the battery life beyond 8 hours in most cases. While the data it produces is useful, we may not use the Microsoft Band 2 in the future because Microsoft recently announced that it discontinued this device.

Samsung Note 7 Recall

The Samsung Note 7 was one of the more popular phones among the original set of participants selected in August. Among 5 participants having Samsung Note 7 devices, all decided to return their devices due to the recall. This discontinued participation for those participants and led to a second round of recruiting in September and an additional month delay.

Data Delivery Format

The collected data are delivered in batches with the format described below. Every file is included in a directory structure with a directory name that describes the data within.

Directory Naming

The data have been delivered in several formats: Apache Avro, CSV, and JSON. Each directory in the delivered data describes the data format, the participant, and the type of data included as follows:

<type>-<participant ID>-<data type>/

For example, directory AVRO-User25-Base/ stores the base data (see below) for the participant coded as User25 in Avro format, while JSON-User10-Label/ stores the label data (see below) for the participant coded as User10 in JSON format.

The following are the directory data types:

Type	Description
Label	Label data type directories include the begin/end label records for interval labeling. This includes begin-end labels for each of the activities listed in Table 9.
Mobility	Mobility data type directories include data related to the radios, connectivity or other types of mobility. This includes: cellular network data, WLAN scans, connectivity, connection strength, and phone state.
Social	Social data type directories include data on social and communication activities. This includes: SMS message logs and call logs.
Survey	Survey data type directories include the results of Lockscreen surveys as well as metadata for interval labels (e.g., placement of phone, placement of watch). This includes results from all Lockscreen surveys from Table 10.
Base	Base data type directories include most of the high frequency sensor and system data. This includes all other sensor and system data from the smartphone or smartwatch (e.g., accelerometer, heart rate, magnetometer, pressure)
<SPECIAL>	<SPECIAL> data type directories are only available for Backers that contributed at the "Choose your own label" level. They contain labels and survey data that correspond to a particular activity or phenomenon of interest.

File Naming

Files within directories are named according to their type and creation time as described below. The basic file naming format is as follows:

<prefix>-<type>-DDMMYYYY-HHMMSS-millis.<format suffix>

Where <prefix> is 'l' for labels, 's' for surveys and metadata, and 'c' for all other types; <type> is the type of data contained (e.g., magnetometer, Riding Ferry, proximity); the date and time are specified numerically using the 24-hour clock; and the <format suffix> is .log for Avro, .json for JSON, and .csv for CSV.

File formats

Each data format presents a different file format. Avro data files contain a header with the schema as well as some basic metadata, while the body of the file contains a list of Avro records. JSON data files contain a list of JSON records delimited by newline characters. CSV files contain rows of comma-separated fields that are separated by newline characters and include a header row.

References

[1] Ferreira, D. et al. AWARE: Mobile Context Instrumentation Framework. Frontiers in ICT, 2015.

[2] Laurila, J. K., The Mobile Data Challenge: Big Data for Mobile Computing Research, MDC, Pervasive 2012

[3] Lohr, S. "CrowdSignals Aims to Create a Marketplace for Smartphone Sensor Data." The New York Times, 22 Mar. 2016 http://nyti.ms/1RwE3Gd

[4] Russel, J. A. A Circumplex Model of Affect. Journal of Personality and Social Psychology, 1980, Vol. 39, No. 6

[5] Schwab et al., Personal Data: The Emergence of a New Asset Class. World Economic Forum Report, 2001

[6] SN Nambi, A. U., et al. LocED: Location-aware Energy Disaggregation Framework. BuildSys 2015

[7] Stisen, A. et al. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. SenSys 2015, pp 127-140

[8] Tangmunarunkit, H. et al. Ohmage: A General and Extensible End-to-End Participatory Sensing Platform. ACM TIST, Vol 6 Issue 3, May 2015

[9] Vaish, R., et al. Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time. CHI 2014.

[10] Welbourne, E. and Cole, G. CrowdSignals.io: Building a Community Dataset. http://igg.me/at/crowdsignals Accessed Oct. 15 2016.

[11] Welbourne, E. and Cole, G. CrowdSignals.io: A Massive New Mobile Data Collection Campaign. http://crowdsignals.io, Accessed Oct 15 2016.

[12] Welbourne, E. and Tapia, E. CrowdSignals: A Call to Crowdfund the Community’s Largest Mobile Dataset. HASCA, Ubicomp 2014

[13] Welbourne, E. et al. Crowdsourced Mobile Data Collection: Lessons Learned from a New Study Methodology. HotMobile 2014.

[14] Wheat RE, et al. Raising money for scientific research through crowdfunding. Trends in ecology & evolution