ALGOSNAP

CrowdSignals.io

Ethics and Institutional Review Board

User's Reference Document

3/01/2016

EMAIL

WEB

111 S Jackson S, Seattle, WA 98104

info@AlgoSnap.com

www.AlgoSnap.com

CrowdSignals.io: Background and Description

Background

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. Personal device data is increasingly one of the largest drivers of innovation in areas as diverse as mobile computing, health, Internet of Things, geography, journalism, marketing, and social science among others. Indeed, personal data is often referred to as "The New Oil" for the 21st century – and several pioneering projects in the last decade (e.g., Nokia Mobile Data Challenge) have demonstrated that precise data, collected ethically from hundreds of smart device owners can supply pertinent, never-before-available information on the specific behaviors, patterns, and trends of individuals, groups, and our society as a whole.

The Problem

Despite this intense focus, there's a critical problem for academic research: smartphone and smartwatch data collection campaigns are extremely expensive, time consuming, and challenging both technically and legally. In addition, many institutions lack the time, funding, and human resources to collect high-quality labeled data from a diverse population. As a result, these data sets 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, recruiting volunteers, administration, and management.

CrowdSignals.io: Creating a Dataset for the Community

CrowdSignals.io is an initiative that begins to address the data problem by crowdfunding a team of experts to collect rich, high-quality smartphone and smartwatch data from a diverse group of 500-1500 participants for 1+ months. Crowdfunding with support from hundreds or thousands of researchers and businesses will ensure that the cost of the dataset is orders of magnitude less than an in-house data collection. By collecting over 50 types of sensor, social, system, and user interaction data, CrowdSignals.io will generate a 50+ TB dataset that is widely applicable to research in many fields. As such, the CrowdSignals.io data will comprise the largest dataset that is accessible to the researcher community. A summary of parameters for the data collection campaign follow:

Participants	500-1500 demographically diverse (e.g., age, geographic location, sex) participants in the United States
Devices	Android Smartphones and smartwatches
Data Types	50+ types of sensor, social, system, and user interaction data, see list below
Ground Truth (Scripted Behavior)	Participants will perform a set of scripted behaviors (e.g., walking, running, eating, cycling) and provide Interval Labels: precise, start-end intervals labeled with specific activities, events, or situations
Ground Truth (Naturalistic Behavior)	For longitudinal, non-scripted behavior, participants will be asked to periodically provide ground truth Point Labels: individual timestamps labeled with specific and ongoing activities, events, or situations
Privacy and Security	We'll use best practices for securing and transporting the data as well as robust, state-of-the-art anonymization techniques (e.g., hashing, feature extraction) for sensitive data types.
Ethics	We'll use interfaces that support data transparency and control for volunteers, and apply best practices for data collection (e.g., informed consent), all documented and available in an "IRB Kit" for researchers.

ALGOSNAP

Ethics and Institutional Review Board Reference

About this Document

The CrowdSignals.io dataset collected using AlgoSnap CrowdSignals Platform will be used by research groups in Universities and Companies around the world. This document is intended as a reference and FAQ for those groups to use in obtaining approval to use the data from their local Ethics Review Committee or Institutional Review Board. As such, the follow sections describe *how the data is collected and how it is managed and protected by AlgoSnap Inc.* Specifically, this document describes: the data collection and storage protocols and the measures taken to protect the confidentiality and anonymity of all human subjects involved. It also documents the CrowdSignals.io data collection methodology for those interested in knowing more about CrowdSignals.io.

AlgoSnap Inc.

CrowdSignals.io is organized and run by AlgoSnap Inc., a Delaware C corporation based in Seattle, WA, USA and founded in 2015. AlgoSnap Inc. is the legal entity responsible for CrowdSignals.io and to which all academic and industry data licensing agreements will refer. AlgoSnap Inc. is also legally the owner of the data so that it can best protect the data and enforce data license agreements. This is because institutions gaining access to the data will have to sign a legal agreement in which, for example, they agree to not attempt to reverse engineer the identities of the data collection participants and to not share the dataset with third parties among other clauses. In addition to the CrowdSignals.io dataset, AlgoSnap Inc. offers services for custom data collections as well as design and evaluation of intelligent algorithms for IoT devices such as smartphones, smartwatches, wearables among others.

Who is Responsible for CrowdSignals.io?

CrowdSignals.io was conceived by and is organized and run by AlgoSnap Inc. Founder and CEO Dr. Evan Welbourne receives expert advising on CrowdSignals.io from leading experts including Prof. Daniel Gatica-Perez (EPFL), Prof. Deborah Estrin (Cornell Tech), and Dr. Henry Tirri (Aalto University).

Evan Welbourne, CEO of AlgoSnap Inc., holds M.S. and Ph.D. degrees from University of Washington, CSE. Since the early 2000s he has worked at the intersection of context-awareness, privacy, and data management for mobile, ubiquitous, and sensor systems. He led the Device Intelligence Group at Samsung Research where his team applied core expertise in context, machine learning, mobility, and predictive analytics to deliver intelligent, personalized user experiences for Samsung's Android and Tizen products. As a Senior Researcher at Nokia he developed the Symbian, Meego, and Android clients for the Simple Context system which was used in the Lausanne Data Collection Campaign – one of the largest longitudinal mobile data collection campaign to date. He has published 20+ papers and holds 15+ patents. He previously managed the Computer Vision Research Group at Amazon, and worked for Microsoft Research and Intel Research.

CrowdSignals.io Methodology Overview

Data Collection System Architecture and Components

CrowdSignals.io uses the AlgoSnap CrowdSignals Data Collection Platform that consists of three components: 1) an app for smartphones and smartwatches that captures, anonymizes, and securely uploads data, 2) a cloud server for securely receiving and storing data, and 3) a website for enrolling and onboarding human subjects as well as for allowing them to delete portions of their uploaded data if necessary.

Once recruited, subjects download and install the app on their personal smartphone and possibly smartwatch. Once installed, the app generates a unique ID (e.g., "138") that serves to anonymously identifies the subject. All data uploaded by a subject's devices will be associated both with this unique ID and a unique device ID that identifies which devices collected the data. The device ID will be mapped to the device's IMEI, but anonymized so that carrier databases cannot be used to reverse engineer subject identities. The app runs in the background on the subject's devices, continuously capturing, anonymizing, and storing sensor, social, system, and user interaction data (see Appendix A: Data Types and Anonymization). The app interface also supports short survey administration and ground truth labeling by subjects; this type of subject feedback can be initiated by the user or triggered automatically by the app (e.g., as in the Experience Sampling Method). When the smartphone is connected to WLAN, the devices periodically securely upload any collected data to the cloud server after which it is deleted from the device. The app applies the latest energy efficient sampling algorithms to ensure that devices will last an entire day of use (e.g., 18+ hours) without needing a battery recharge.

The cloud server leverages commercial cloud services (e.g., Google Cloud Platform) to securely and scalably receive and store the data uploaded by subjects' client devices. Data on the server is indexed and stored in a database for efficient management by AlgoSnap Inc. and potentially by subjects themselves via the website. The server also applies algorithms to validate the quality and integrity of the collected data – and it may apply further anonymization algorithms before the data is finally shared with researchers.

The website provides the primary interface through which data collection administrators (i.e., AlgoSnap Inc.) interact with subjects, and through which subjects understand and manage the data they have uploaded. The website includes several interfaces and procedural flows that are critical to the CrowdSignals.io data collection process. The first such procedural flow is recruiting and onboarding, which is supported by an interface that educates potential subjects on the benefits and risks of the research being conducted and walks them through the informed consent process (including legal documents). This process is personally assisted by AlgoSnap Inc. administrators who may be in direct contact with potential subjects via audio, video, or text conferencing software. A second procedural flow is supported by a series of web pages that walks subjects through the process of creating an account, receiving a unique CrowdSignals.io subject ID (e.g., "138"), completing an initial demographic survey, and downloading and installing the app to their smartphone and smartwatch. A third interface allows subjects to log in and review, correct, or delete the data they have already uploaded. This interface shows detailed information and visualizations on each data type and file uploaded over time.

Subject Recruitment, On-boarding, Management, and Compensation

Subjects are recruited through web platforms and email-based messages that advertise the opportunity to participate in a paid smart device data collection study. Interested people may respond to the email to further discuss the opportunity with AlgoSnap Inc. administrators who will share a link to the website where the subject can complete the informed consent and on-boarding process. AlgoSnap Inc. administrators are constantly in touch and available to subjects throughout the data collection process to answer questions and address any technical or other difficulties. After the data collection period has ended, subjects will receive compensation

via the website infrastructure (e.g., PayPal) comprised of a base pay amount and an additional payment pro-rated by the amount of data and feedback or event labels the subject provided during the study.

Researcher Qualifications

Who are the principal and associated investigators?

The investigation is led by AlgoSnap, as such the principal investigator is Evan Welbourne, PhD (UW), Chief Executive Officer of AlgoSnap Inc.

Have investigators completed human subjects training at an institution?

Yes, Evan Welbourne had human subjects training at the University of Washington (UW)

What are the researchers' qualifications to conduct this study?

Evan Welbourne, PhD

Evan has designed, supervised, and administered research studies involving human subjects since 2001 at the University of Washington and in at Intel Research, Nokia Research, and Samsung Research. Key examples in academia include the leadership of the RFID Ecosystem project at the University of Washington, CSE – a large-scale, longitudinal data collection project that tracked thousands of people and objects using passive RFID, as well as a longitudinal study of mobility patterns in mobile healthcare workers using a multimodal sensing platform. User studies in industry included work on the Lausanne Data Collection Campaign at Nokia and crowdsourced data collection campaigns at Samsung.

Short Bio: Evan, CEO of AlgoSnap Inc., holds M.S. and Ph.D. degrees from University of Washington, CSE. Since the early 2000s he has worked at the intersection of context-awareness, privacy, and data management for mobile, ubiquitous, and sensor systems. He led the Device Intelligence Group at Samsung Research where his team applied core expertise in context, machine learning, mobility, and predictive analytics to deliver intelligent, personalized user experiences for Samsung's products. As a Sr. Researcher at Nokia he developed the Symbian, Meego, and Android clients for the Simple Context system which was used in the Lausanne Data Collection Campaign – one of the largest longitudinal mobile data collection campaign to date. He has published 20+ papers and holds 15+ patents. He previously managed the Computer Vision Research Group at Amazon, and worked for Microsoft Research and Intel Research.

Study Description

What is the title of the study?

CrowdSignals.io: Building the Community's Largest Labeled Mobile and Sensor Dataset.

Where is the research going to take place?

The data collection is taking place across all states of the United States of America.

How much time will be needed to conduct and complete the research study?

The data collection will take three months to complete, out of which subjects will be asked to collect data between one and two months.

What is the source of the funding for this research?

Funds collected from an online crowdfunding campaign as well as sponsorship from several large organizations.

What are the anticipated dates of research?

The data collection will happen between February and May 2016.

Where will the study be carried out (e.g. public place, in researcher's office, in private office at organization)?

The data collection will happen at any place recruited subjects visit during the normal course of their lives. Participant subjects will install a data collection app in their smartphone and smartwatch that will collect the subject's data as they go about their life.

What is the purpose of the study?

The purpose is to create a unique massive mobile and sensor dataset by performing an ethical data collection from thousands of subjects across the United States owning Android smartphone and smartwatch devices. The data collected will include 50+ types of sensor, social, system, and interaction data from these devices. This massive dataset will enable researchers around the world to solve critical problems for which there is presently a lack of data. In the process, participant subjects will be educated on the data their devices generate while informing them of the risks and benefits of sharing it.

What is it that you hope to learn from the study and the assessed importance of this new knowledge?

The goal is to collect a unique and massive dataset that will enable researchers around the world to solve critical problems for which there is presently a lack of data. The benefits for society at large could be enormous by allowing scientists worldwide to solve problems in areas as diverse as computing, journalism, public health, social science, and urban planning among others.

What is the detailed study protocol or procedure? Describe ALL the procedures human subjects will undergo. Are the research procedures the least risky that can be performed consistent with sound research design.

For the CrowdSignals.io data collection study, a diverse group of 500-1500 subjects will be recruited for 1+ months across the United States to provide the data collected from their Android smartphone and smartwatches.

A summary of parameters for the data collection campaign are as follow:

Participants	500-1000+ demographically diverse (e.g., age, geographic location, sex) participants in the United States
Devices	Android Smartphones and smartwatches
Data Types	50+ types of sensor, social, system, and user interaction data, see list below
Ground Truth (Scripted Behavior)	Participants will perform a set of scripted behaviors (e.g., walking, running, eating, cycling) and provide Interval Labels: precise, start-end intervals labeled with specific activities, events, or situations
Ground Truth (Naturalistic Behavior)	For longitudinal, non-scripted behavior, participants will be asked to periodically provide ground truth Point Labels: individual timestamps labeled with specific and ongoing activities, events, or situations
Privacy and Security	We'll use best practices for securing and transporting the data as well as robust, state-of-the-art anonymization techniques (e.g., hashing, feature extraction) for sensitive data types.
Ethics	We'll use interfaces that support data transparency and control for volunteers, and apply best practices for data collection (e.g., informed consent), all documented and available in an "IRB Kit" for researchers.

The CrowdSignals.io data collection platform consists of three components: 1) an app for smartphones and smartwatches that captures, anonymizes, and securely uploads data, 2) a cloud server for securely receiving and storing data, and 3) a website for enrolling, onboarding human subjects and allowing them to delete information and data collected if necessary.

Once recruited, subjects download and install the app on their personal smartphone and possibly smartwatch as well. Once installed, the app generates a unique ID (e.g., "138") that will serve as an anonymous identifier for the subject. All data uploaded by a subject's devices will be associated both with this unique ID and a unique device ID that identifies which devices collected the data. The device ID will be mapped to the device's IMEI, but anonymized so that carrier databases cannot be used to reverse engineer subject identities. The app runs in the background on the subject's devices, continuously capturing, anonymizing, and storing sensor, social, system, and user interaction data (see Appendix A: Data Types and Anonymization). The app interface also supports short survey administration and ground truth labeling by subjects; this type of subject feedback can be initiated by the user or triggered automatically by the app (e.g., as in the Experience Sampling Method). When the smartphone is connected to WLAN, the devices periodically securely upload any collected data to the cloud server after which it is deleted from the device. The app applies the latest energy efficient sampling algorithms to ensure that devices will last an entire day of use (e.g., 18+ hours) without needing a battery recharge.

The cloud server leverages commercial cloud services (e.g., Google Cloud Platform) to securely and scalably receive and store the data uploaded by subjects' client devices. Data on the server is indexed and stored in a database for efficient management by AlgoSnap Inc. and potentially by subjects themselves via the website if they decide to delete a specific portion of the data uploaded. The server also applies algorithms to validate the quality and integrity of the collected data – and it may apply further anonymization algorithms before the data is finally shared with researchers.

The website provides the primary interface through which data collection administrators (i.e., AlgoSnap Inc.) interact with subjects, and through which subjects understand and manage the data they have uploaded. The website includes several interfaces and procedural flows that are critical to the CrowdSignals.io data collection process. The first such procedural flow is recruiting and onboarding, which is supported by an interface that educates potential subjects on the benefits and risks of the research being conducted and walks them through the informed consent process (including legal documents). This process is personally assisted by AlgoSnap Inc. administrators who may be in direct contact with potential subjects via audio, video, or text conferencing software. A second procedural flow is supported by a series of web pages that walks subjects through the process of creating an account, receiving a unique CrowdSignals.io subject ID (e.g., “138”), completing an initial demographic survey, and downloading and installing the app to their smartphone and smartwatch. A third interface allows subjects to log in and review, correct, or delete the data they have already uploaded. This interface shows detailed information and visualizations on each data type and file uploaded over time.

What led to the formulation of the study?

There are presently many critical problems that researchers cannot solve due to the lack of data combined with the great challenge and expense of collecting it. This is an effort to collect the data many researchers need to advance the state-of-the-art in many fields of study.

Drugs, Devices and Substances

Will drugs, placebos, biological agents or other substances (such as food substances or vitamins) be administered as part of this study?

No

Will any invasive or potentially harmful procedures of any kind will be used?

No

Will an investigational medical device be used?

No

Will radiation or radioactive materials be used?

No

Will special diets be used?

No

Will the research study involve storage and/or analysis of human tissue or any human materials?

No

Will your study involve working with any substances and / or equipment, which may be considered hazardous?

No, subjects will not be explicitly requested to perform any activities or interact with any equipment that is not normally part of their daily routine, work, employment, or habits or do not know how to operate. Data will be simply collected in the background from subjects as they go about their lives. When subjects are specifically requested to provide labeled examples of an activity or situation, subjects will be asked to perform it only if they are comfortable doing it, if it is part of their normal routine, and if they know how to do it without incurring additional risks to their physical, psychological, or social wellbeing.

Are alcoholic drinks to be administered to the study subjects?

No, not explicitly requested. However, since data is collected form subjects in the background as they go about their normal lives subjects might consume alcoholic beverages during the data collection depending on their daily normal routine.

Anything else that might be potentially harmful to subjects in this research?

No. Since data is collected in the background while subjects go about their life, the only harmful/risky activities are those normally or routinely performed by the subjects. Also, when subjects are specifically requested to provide labeled examples of an activity or situation, subjects will be asked to perform it only if they are comfortable doing it, if it is part of their normal routine, and if they know how to do it without incurring additional risks to their physical, psychological, or social wellbeing.

Subject Population

How many human subjects will be included in the study?

The data collection study will include between 500 and 1500 participant subjects.

What will be the demographics of the subjects in the study (age range, gender, race or ethnicity, etc.)?

Data will be collected from participants at least 18 years of age and older. Efforts will be made to recruit participants from diverse demographics within these age limits.

What is the total time subjects will spend in the entire study (e.g. 20mins, 3hours, 4 days, etc.)?

Subjects will participate in the data collection study between one and two months.

How will you recruit participants (e.g. ads, word of mouth, letters mailed home, email, etc.)?

Subjects from across the United States will be recruited via online web platforms and email.

Where will subjects be recruited and located during the study?

Subjects will be recruited and located in their usual area of residency or travel in the United States during the data collection study.

Who will approach potential participants?

Subjects will be approached via online platforms and email by AlgoSnap Inc., the organizer of this data collection study.

What is the inclusion/exclusion selection criteria for the human subjects in the study?

The following is the subject's inclusion criteria:

- At least eighteen (18) years old or older.
- A resident of the United States and physically reside in the United States.
- No chronic medical conditions or disabilities such as impaired decision-making capacity.
- Understand the English language.
- Do not reside in any care facilities such as foster care, elderly care and prisons.
- Not be elected or appointed public officials or candidate for public office as well as members of the media such as media reporter, journalist, columnist, editor, or blogger.
- Own one Android smartphone and use it as subject's primary smartphone and have a monthly voice, and data plan.
- Own an Android smartwatch and use it as subject's primary smartwatch.
- Have installed the AlgoSnap Inc. Android app on said Android smartphone and possibly smartwatch for data collection.
- Use said Android smartphone and potentially smartwatch as subject normally does, between one and two months.
- Answer a demographics questionnaire.
- Answer a questionnaire about subject's and information regarding their smart devices, communication patterns, and frequent locations visited.

- Agree to provide the requested number of activity or situations labels (short surveys) during the data collection study. These short surveys that are administered via their smartphone several times per day. Each smartphone-based survey will take less than 30 seconds and provides “ground truth” information on the subject’s current activity or situation.
- Uninstall the AlgoSnap Inc. application upon withdrawal from participation if applicable.
- Agree to be bound by the terms of this Research Study Participant Agreement, including all Attachments.

All subjects will be fully informed about the study before they consent to participate and before any data is collected.

Will the research study include women, minorities, or minors? Provide a rationale for not including these populations if the research might benefit these groups.

The data collection study will include women and minorities. Minors will not be included to simplify the data collection legal paperwork since parents or guardians would need to provide consent in this case.

Will any participants be your students, laboratory personnel and/or employees?

No, AlgoSnap Inc. personnel will be excluded from the data collection study.

Will your study involve participants who are particularly vulnerable or unable to give informed consent or in a dependent position (e.g. children or people under 18, pregnant woman, impaired decision-making capacity, non-English speakers, homeless, over-researched groups or people in care facilities such as foster care, elderly care and prisons)?

No

What is the inclusion criteria for any vulnerable population and why is that population being used?

Not Applicable

What are examples of advertisements/notices and letters to potential subjects?

Examples will be provided in January 2016 when the CrowdSignals.io pilot data collection is executed.

Describe all plans to compensate subjects in cash or other form of payment (e.g. gift card, lottery, draw ticket, etc.). Describe how payment will be prorated if any.

Subjects will be offered a base payment for their participation in the study, delivered electronically (e.g., via PayPal). Subjects will also be offered additional payment, pro-rated by the amount of data and feedback or event labels they upload.

Will subjects be reimbursed for travel expenses?

No, subjects will not be reimbursed for travel expenses and no additional travel will be requested from subjects participating in this data collection other than that involved in their normal daily routines.

Risk for Researchers

Could the nature or subject of the research potentially expose the researcher(s) to threats of physical violence and / or verbal abuse?

No. Researchers will only interact with participant subjects remotely via online conversations, email, and telephone calls.

Could the nature or subject of the research potentially have an emotionally disturbing impact on the researcher(s)?

No

Will any researchers be in a lone working situation?

No

Anything else that might be potentially harmful to the researcher(s) in this research?

No. Researchers will only interact with participant subjects remotely via online conversations, email, and telephone calls.

Risk for Subjects

What are the risks for subjects participating in this study (physical, psychological, economic, social wellbeing)?

There are no additional risks during this data collection study other than those normally found in subjects' normal daily routine as data is being recorded in the background as subjects go about their lives.

When subjects are specifically requested to provide labeled examples of an activity or situation, subjects will be asked to perform it only if they are comfortable doing it, if it is part of their normal routine, and if they know how to do it without incurring additional risks to their physical, psychological, or social wellbeing.

Could the study induce psychological stress or anxiety, or produce humiliation or cause harm or negative consequences beyond the risks encountered in the everyday life of the subjects?

No

Could disclosure of the subjects' responses outside the research reasonably place them at risk of civil or criminal liability, or be damaging to their financial standing, employability or reputation?

No

Will the study involve discussion of sensitive topics (e.g. sexual activity, drugs, ethnicity, illegal activities)?

No

Are all or any of the subjects either elected or appointed public officials or candidate for public office?

No

What are the risks, discomforts associated with each intervention or procedure in the study?

There are no risks or discomforts associated with this data collection study other than those found in subject's everyday life. When subjects are specifically requested to provide labeled examples of an activity or situation, subjects will be asked to perform it only if they are comfortable doing it, if it is part of their normal routine, and if they know how to do it without incurring additional risks to their physical, psychological, or social wellbeing.

How will you arrange for professional intervention if you believe is necessary?

Not applicable

Is it possible that you will discover a subject's previously unknown condition (disease, suicidal intentions, genetic predisposition, etc.) as a result of study procedures?

No, this is unlikely. Even when data about subject's daily routine will be collected, sensitive data will be anonymized with state-of-the-art techniques to prevent these types of situations.

Are you conducting research outside the United States of America? No

Benefits

What are the potential benefits to be gained by society as a result of this study?

Each researcher or group of researchers working with the dataset collected in CrowdSignals.io will need to respond individually to argue the societal benefits of their particular program of research. Here we provide a general description of how the CrowdSignals.io dataset benefits society.

Personal device data is increasingly one of the largest drivers of innovation in areas as diverse as mobile computing, health, Internet of Things, geography, journalism, marketing, and social science among many others. Indeed, personal data is often referred to as "The New Oil" for the 21st century – and several pioneering projects in the last decade (e.g., Nokia Mobile Data Challenge) have demonstrated that precise data, collected ethically from hundreds of smart device owners can supply pertinent, never-before-available information on the specific behaviors, patterns, and trends of individuals, groups, and our society as a whole. As such, the research enabled by the CrowdSignals.io dataset has the potential to positively impact society in a great variety of areas. Several examples follow:

Field	Research	Benefit to Society
Epidemiology	Study how mobility patterns (e.g., proximity of households to services, use of public transit) impact choices on food purchases.	Fundamentally new and grounded insights on how people make food choices. This information could inform the design of a variety of new civic projects.
Health Tracking	Use of diverse, large scale data on human fitness activities to train robust machine learning models that can recognize and track these activities (e.g., cycling, running, walking).	Robust fitness tracking algorithms could be used by companies and individuals alike to better understand and improve health in our society.
Social Science	Investigation of how social recognizers facilitate assessment and analysis of interpersonal social behavior.	Fundamentally new discoveries in social science could lead to improved interventions that improve our communication and facilitate interaction in a variety of settings.
Human Computer Interaction	Study of the most effective ways to help end-users understand machine learning results over multimodal data.	New techniques that aid end-users in understanding machine learning results could be fundamental to creating safe and effective intelligent systems.

What are the potential benefits to be gained by subjects as a result of this study?

Each researcher or group of researchers working with the dataset collected in CrowdSignals.io will need to respond individually to argue the benefits of their particular program of research to the subjects. Here we provide a general description of how the CrowdSignals.io dataset benefits subjects. First, all subjects will benefit directly from the compensation they receive for their participation. Additional benefits include an education and deeper understanding of their own personal data as well as the risks and benefits of sharing it. In particular, as part of the informed consent process, all subjects will learn about the data generated by their personal devices (smartphone and smartwatch) as well as the risks and benefits of sharing it, and the kinds of information that could be derived from it. Subjects will also have the opportunity to review and study their own personal data by logging into the web interface and accessing logs, visualizations, and summaries of their data. Insights offered by the website include but are not limited to: understanding of personal mobility patterns, physical activity levels over time, actual device and app usage patterns, and patterns in communication and sociability via these devices.

Data Collection

What data will be collected from subjects (e.g., interview, questionnaire, field observation, sensor data)?

A combination of questionnaires and mobile and sensor data will be collected from subjects. On entering the study, subjects will complete a web-based questionnaire that solicits basic information on demographics (e.g., age range, level of education). During the study, subjects will complete a number of brief (e.g., less than 1 minute) questionnaires administered via their smartphone. These brief, in-situ questionnaires gather "ground truth" information from subjects regarding their current activity, situation, or state (e.g., "walking", "having lunch with a friend"). A variety of network, sensor, social, system, location, and interaction data will also be logged by the subject's devices and uploaded to the cloud server. This data comprises the bulk of the CrowdSignals.io dataset and is described in detail below in Appendix A: Data Types and Anonymization.

Where will the data be collected?

The data will be collected throughout the United States.

How will the data be collected?

The data will be collected using the system and processes described in the CrowdSignals.io Methodology Overview above. An abbreviated description of the data collection process is as follows.

On entering the study, subjects will complete a web-based demographic questionnaire. Subjects will also download and install an app to their smartphone and possibly smartwatch. This app will log and upload network, sensor, social, system, location, and user interaction data from the devices for the duration of the data collection period. The app will also periodically administer brief (e.g., less than 1 minute) questionnaires that are designed to collect ground truth information on the subject's current activity, situation, or state (e.g., "walking", "eating lunch with a friend").

Is there audio or video recording?

There will be no audio or video recording as part of the study. However, as part of the sensor data collection by smartphones, degraded audio (i.e., derived audio features) will be collected. The degraded audio will be collected periodically in short bursts (e.g., 4 seconds at a time) and will be designed in such a way so that the original, human-intelligible audio stream is extremely difficult to be reconstructed using state-of-the-art techniques.

Will questionnaires be completed anonymously and returned indirectly?

Questionnaires will be associated with the subject's unique ID (e.g., "138").

Will questionnaires and/or interview transcripts only be identifiable by a unique identifier?

Yes, all data collected in the study including questionnaire data will only be identifiable by a unique ID.

How will the data collected be used (e.g., analytics, reports, publications)?

The data collected will be used by research and development groups around the world to conduct research and advance the state-of-the-art in a wide variety of data-driven areas. Analytics, publications, and software are all possible outcomes of work with the CrowdSignals.io dataset.

Privacy and Confidentiality

Can you explain where the research study takes place and how will you maintain privacy in this setting?

The data collection will take place in the United States, and specifically at all locations within the United States that are occupied by subjects during the data collection period. Since a subject's interaction with the CrowdSignals.io data collection occurs entirely online and via their smart devices, privacy will be maintained by following best practices for secure communication and transmission of data (e.g., user accounts, anonymization, encryption, secure networking).

Will all personal information gathered be treated in strict confidence and never disclosed to any third parties?

All personal information will be treated in strict confidence and never disclosed to any third parties without explicit informed consent from the subjects. All personally identifiable information (i.e., name, address) will be treated in strict confidence and never disclosed to any third parties.

What identifiable data will you obtain from participants?

Subjects' name, address, phone number, and electronic contact information (e.g., email address) will be collected during the on-boarding process to verify the subjects' residence in the United States and to facilitate payment of compensation.

What anonymization procedures and physical and technical security measures will be used to protect the personal data collected?

All subjects will receive a unique ID (e.g., "138"), which will be stored in place of any personally identifiable information. The link between subject ID and the subject's name and address will be stored separately in a secure, offline location. All data collected from smart devices will be anonymized on the device before upload to the cloud server, and in some cases a second time on the cloud server before the dataset is released to researchers. Data will be stored using secure cloud storage systems. More detail on anonymization techniques for sensitive data types gathered from smartphone and smartwatches is provided below in Appendix A: Data Types and Anonymization.

Will data be associated with personal identifiers or will it be coded? Will it be possible to link personal data back to individual subjects in any way? (This includes linking identities but does not include identifying participants from signed consent forms that are stored securely separate from their research data)

The data will be associated with unique identifiers for each subject. The personal data will not be linked back to individual subjects except by the unique identifier – and the identifier-subject mapping will be stored separately offline.

If you plan to code the data, describe the method in which it will be coded and indicate who will have access to the key to the code.

The data will be coded so a unique ID instead of personal identifiable information represents subjects' data. The data will be also post-processed for anonymization as described below in Appendix A: Data Types and Anonymization. For us, coding means replacing personal identifiable information with a unique subject ID that cannot be traced back to the personal information.

Will all place names and institutions which could lead to the identification of individuals or organizations be changed?

All names of places and institutions which could lead to the identification of individuals will be changed or at least made sufficiently generic (e.g., “Whole Foods Market Union Square” → “Whole Foods” or “Whole Foods” → “Grocery”) as to strengthen anonymity.

Will lists of identity numbers or pseudonyms linked to names and/or addresses be stored securely and separately from the research data?

Yes, the mapping from ID numbers to names and addresses of subjects will be stored separately and securely.

Will all personal information related to this study be retained and shared in a form that is fully anonymized?

Yes, all personally identifiable information (e.g., name, phone, address, email) will be replaced with a unique identifier. All other personal data collected from subjects will be retained and shared in a form that is anonymized.

Where will the data be stored and how will it be secured? How will you ensure the security of identifiable data (password protected computer, encrypted files, locked cabinet, locked office)?

Researchers using the CrowdSignals.io dataset must answer this question individually and in compliance with the data license, the following answer describes the approach taken by AlgoSnap Inc.

The anonymized personal data will be stored on a cloud server (e.g., Google Cloud Platform) that is secured through means provided by the platform as well as standard protections including access control and password protection.

Describe how identifiable data will be transferred (courier, email) or transmitted (e.g. file transfer software, file sharing, email)?

At the time of consent and on-boarding, a subject’s name, phone number, address, and email will be transmitted via a secure web interface or via a phone call. During data collection, all collected data will be transmitted to the cloud server from smart devices via secure http connection. When data is shared with a researcher or research group, it will be shared securely via an account on a cloud server platform (e.g., a Google Drive).

Are the data, documents, and/or records publicly available?

No, none of the data or documents is publicly available.

Do you plan to use or disclose identifiable health information (HIPAA privacy rule)?

No, no identifiable health information will be collected or disclosed.

Consent and Information

Describe in detail your consent methods, process, and settings. Identify who will provide the information to subjects and who will interact with them during the consent process.

The consent process will take place online via a web interface, assisted by an AlgoSnap Inc. administrator via electronic messaging and possibly voice or text conference. The consent process will proceed as follows: First, potential subjects are invited to participate by web platforms or email advertisement. Before deciding whether or not to consent, subjects that respond to the advertisement must read through extensive educational materials including plain English descriptions and simple diagrams that describe the CrowdSignals.io data collection purpose and process as well as the risks and benefits of participating. Subjects will also have the opportunity have any questions answered by AlgoSnap Inc. administrative staff before deciding whether or not to participate.

Will all subjects be given an Information sheet and be given adequate time to read it and ask questions before being asked to agree to participate?

Yes, because the educational portion of the consent process may happen asynchronously, subjects will have ample time (e.g., hours, days) to read about the study and ask questions before being asked to consent.

Will all subjects taking part of this study be asked to sign a fully informed consent form freely (e.g. if participating in an interview, focus group, observation, data collection, etc.)? If you are obtaining consent another way, please explain under.

Yes, all subjects taking part in this study will be asked to sign a fully informed consent form freely.

What are examples of the informed consent forms that subjects will sign?

Examples will be provided following the CrowdSignals.io pilot study in early 2016.

When and how will subjects be requested to provide informed consent?

Subjects will be requested to provide consent by AlgoSnap Inc. administrative staff via electronic message after they have voluntarily responded to the invitation to participate and completed the educational portion of the informed consent process.

What procedures will you use to assess if the subject understands the information contained in the consent?

During the educational portion of the informed consent process, subjects will be required to take several brief knowledge tests to assess their understanding. AlgoSnap Inc. administrative staff will also discuss the study and the educational material with subjects to assess their understanding and answer questions before asking them to consent.

Are you planning to enroll subjects who do not have the capacity to consent?

No, only subjects that have the capacity to consent will be enrolled.

Will all subjects be told that they can withdraw at any time, ask for their data to be destroyed and/or removed from the study and when this will no longer be possible (e.g. once it has been included in the final report)?

Yes, all subjects will be told that they can withdraw at any time and that they may ask for their data to be destroyed and/or removed from the dataset. Moreover, subjects may directly withdraw and delete their own data using the website at any time. Subjects may also use the app interface to turn off all data logging at any time and for any length of time during the data collection period. However,

subjects will not be paid for periods when data collection was turned off. All subjects will be informed as to the end of the data collection period after which their data may no longer be deleted.

What are the exact arrangements for withdrawal of participation and withdrawal of data from your study?

Subjects may withdraw from the data collection at any time during the data collection period. At this time, they may request that all data collected from them be destroyed, subjects may also log in to the website to delete any or all of their data at any time. To withdraw from the study and request that their data be destroyed, subjects simply need to send an email to an AlgoSnap Inc. administrator within the data collection period. Subjects who withdraw from the data collection and request that their data be destroyed will not receive full payment for their participation due to early termination.

Will all subjects self-completing a questionnaire be informed that returning the completed questionnaire implies consent to participate?

Subjects will not complete any questionnaires until they formally agree to consent and sign a consent form.

Will participants be fully informed about the purpose of the study?

Yes, subjects will be fully informed about the purpose of the study during the educational portion of the informed consent process.

Will subjects be required to take part in the study when information about the research purpose and design is withheld from them? (e.g. covert observation of people in non-public places), and / or will deception of any sort be used)

No deception of any sort will be used and no part of the research purpose and design will be withheld from subjects.

Can data acquired in the study affect a subject's relationship with other individuals (e.g. employee-supervisor, patient-physician, student-teacher, family relationships)?

Since the collected data is of a personal nature (e.g., patterns of communication, physical activity, mobility) it is conceivable that the data may affect a subject's relationship with other individuals. However, this risk is minimized by the fact that the subject's data will be anonymized and by the fact that no individuals other than the subject and the researchers will have access to the collected data.

Describe how you will minimize any undue influence on your subjects' decision about participating in your research.

All subjects will respond voluntarily to an email invitation from AlgoSnap Inc. Neither the email nor any part of the consent process will involve persons with authority over a subject such as the subject's teacher, doctor, or employer.

What follow-up efforts will be made to detect any harm to subjects?

AlgoSnap Inc. will remain in contact will subjects for the duration of all data licenses and will directly receive any instances of harm reported by subjects themselves. In addition, the collected data will be individually watermarked for each group it is shared with so that any instances of data leakage or abuse can be traced.

Do you expect that all of your subjects will be fluent in spoken and/or written English?

Yes, all subjects will be fluent in spoken and/or written English.

Data Management and Responsibilities

State how long study information including research data, consent forms and administrative records will be retained, in what format(s) and where the information will be kept.

The study information including collected data, consent forms, and administrative records will be retained and stored electronically on a cloud server or in secure local storage for the duration of the data licenses.

What will happen to the data when the study is completed?

The data will be retained for the duration of the data license and then it must be destroyed by the different institutions having granted access to the data via a data license.

[Europe?] Will you ensure that the processing of personal information and personal identifiable information related to the study will be in full compliance with the Data Protection Act 1998 (DPA)?

Researchers in Europe must take steps to comply.

[Europe?] Will processing any personal information outside of the European Economic Area (EEA) take place and if so, how compliance with the DPA will be ensured?

Researchers can use information provided in this document to write about compliance with DPA regarding processing of personal information for subjects in the United States.

Will the Principal Investigator take full responsibility during the study, for ensuring appropriate storage and security of information (including research data, consent forms and administrative records) and, where appropriate, will the necessary arrangements be made in order to process copyright material lawfully?

Researchers must take full responsibility for ensuring the appropriate storage and security of the CrowdSignals.io dataset for the duration of their data license with AlgoSnap Inc. This will be enforced by a legal agreement before researchers are granted access to the data collected in CrowdSignals.io.

AlgoSnap Inc. will take full responsibility for ensuring appropriate storage and security of information during the data collection period and for the storage and security of the master copy of the information for the duration of all data licenses.

Who will have access to personal information relating to this study?

Only AlgoSnap Inc. will have access to personally identifiable information (i.e., names, addresses) of subjects in the CrowdSignals.io dataset.

Will you use the research data for any purpose other than that which consent is given?

No, research data will only be used for the purpose for which consent is given.