## Background

Mobile sensory data enables the research and development of algorithms and solutions to many problems. The dataset provided here contains data from 60 users (also referred to as subjects or participants), each identified with a universally unique identifier (UUID). From every user, it has thousands of examples, typically taken in intervals of 1 minute (but not necessarily in one long sequence, as there are time gaps). Every example contains measurements from sensors (from the user's personal smartphone and from a smartwatch that we provided). Most examples also have context labels self-reported by the user.

Dataset URL: [Google Drive](https://drive.google.com/file/d/1-AyizDvWkVVsp_NjCYXp_h5PbbPVlQ7B/view?usp=sharing)

## How data was collected

Data was collected using a mobile application having both a version for iPhone and for Android, with a Pebble watch component that interfaces with both the iPhone and the Android versions. The app performs a 20-second "recording session" automatically every minute. In every recording session, the app collects measurements from the phone's sensors and from the watch (if it is available), including:

* the phone's accelerometer, gyroscope, and magnetometer (sampled in 40Hz),
* audio (sampled in 22kHz, then processed to MFCC feature representation),
* location, the watch's accelerometer (sampled in 25Hz) and compass,
* and additional sensors, if available (light, humidity, air pressure, temperature).

The measurements from a recording session are bundled into a zip file and sent to the lab's web server (if WiFi is available or stored on the phone until WiFi is available).

## Users

The users were primarily students (both undergraduate and graduate) and research assistants from the UCSD campus.

* 34 iPhone users, and 26 Android users.
* 34 female, 26 male.
* 56 right-handed, 2 left-handed, and 2 defined themselves as using both.
* Diverse ethnic backgrounds (each user defined their "ethnicity" how they liked), including Indian, Chinese, Mexican, Caucasian, Filipino, African American, and more.

## Sensors

The sensors used were diverse and included high-frequency motion-reactive sensors (accelerometer, gyroscope, magnetometer, watch accelerometer), location services, audio, watch compass, phone state indicators, and additional sensors that were sampled in low frequency (once a minute).

Not all sensors were available all the time. Some phones didn't have some sensors (e.g. iPhones didn't have an air pressure sensor). In other cases, sensors were sometimes unavailable (e.g. location services were sometimes turned off by the user's choice, and audio was not available when the user was on a phone call).

The following table specifies the different sensors, the format of their measurements for a single example, and the total number of labeled examples (#ex) and users (#us) that have measurements from each sensor.

| **sensor** | **details** | **dimension** | **#us** | **#ex** |
| --- | --- | --- | --- | --- |
| accelerometer | Tri-axial direction and magnitude of acceleration. 40Hz for ~20sec. | (~800) x 3 | 60 | 308,306 |
| gyroscope | Rate of rotation around phone's 3 axes. 40Hz for ~20sec. | (~800) x 3 | 57 | 291,883 |
| magnetometer | Tri-axial direction and magnitude of magnetic field. 40Hz for ~20sec. | (~800) x 3 | 58 | 282,527 |
| watch accelerometer | Tri-axial acceleration from the watch. 25Hz for ~20sec. | (~500) x 3 | 56 | 210,716 |
| watch compass | Watch heading (degrees). nC samples (whenever changes in 1deg). | nC x 1 | 53 | 126,781 |
| location | Latitude, longitude, altitude, speed, accuracies. nL samples (whenever changed enough). | nL x 6 | 58 | 273,737 |
| location (quick) | Quick location-variability features (no absolute coordinates) calculated on the phone. | 1 x 6 | 58 | 263,899 |
| audio | 22kHz for ~20sec. Then 13 MFCC features from half overlapping 96msec frames. | (~430) x 13 | 60 | 302,177 |
| audio magnitude | Max absolute value of recorded audio, before it was normalized. | 1 | 60 | 308,877 |
| phone state | App status, battery state, WiFi availability, on the phone, time-of-day. | 5 discrete | 60 | 308,320 |
| additional | Light, air pressure, humidity, temperature, proximity. If available sampled once in session. | 5 | --- | --- |

## Understanding the Data

For information on the data feature, see the attached README.md

## Goal

The goal of this project is as follows:

1. Each user has a separate data file that contains several pieces of the data being features and labels. Write some code to read in the required data for subsequent processing and analysis.
2. Explore the data and provide some insights on the following points:
   * what labels are included in the data? How do the reported contexts change through time? What combinations of labels are common?
   * Which sensors were captured in the recorded data?
   * The relationship between features and labels
3. Build a model for predicting the context label (the activity of a user) based on the sensor-features. Your choice of model is left to your discretion

## Deliverable

Your submitted solution pack should include the following:

* A python source code (Jupyter Notebook) for the entire analysis submitted via GitHub Repository
* A short PowerPoint presentation no more than 5 content pages (plus 1 page for title, and one for appendix/references, totalling 7 pages in all) summarizing the result of your work (which you may/may not present orally)
* Submit your project using this [Google Form](https://forms.gle/ihmwU3Rgf6AdZTAH7)
* Key references used in the completion of the project should be listed in the project report

## Due Date

The challenge begins from today 1st December, 2022 at 8:00 AM till Monday 5th December, 2022 at 12:00 Noon. All participants must adhere to these timelines for our collective effectiveness.

## Honor Code

No collaboration is expected between teammates, however participants may interact with teammates as part of general research for completing the work at hand.

## Evaluation criteria

**Data quality assessment (5%)**

Assess the data provided and take notes of any data quality issues present. If available provide suggestions of how this might be addressed. If no obvious data quality issue is present, take note and indicate this as well in your report

**Quality of data insights (40%)**

Provide any general initial insight about the data through visualizations and techniques as needed.

**Model Building (50%)**

Build a model for the purpose of predicting the context label given the sensor data.

**Presentation (5%)**

Overall quality of the work presentation