# Bar Crawl: Detecting Heavy Drinking: An Analysis of Accelerometer Data and its Ability to Predict TAC

Aidan Simpson, Alyssa Ngyuen, Audrey Kimball, Connar Gibbon

## Overview

Our project aims to look at and analyze the behavioral patterns of heavy drinking events, like a college bar crawl, our data set from UCI machine learning repository, *Bar Crawl: Detecting Heavy Drinking*. The data set provides multiple types of data points that can be used to detect and understand factors that predict heavy drinking behavior in social settings. The motivation of the study comes from the public health importance of noticing and addressing heavy drinking behaviors. With the growing concerns about social and health consequences of excessive drinking, our research could help with the development of strategies for monitoring and potentially stopping risky drinking behaviors. Analysis of this data we are hoping to identify patterns that could create interventions about targeting high-risk drinking situations.

## Description of Dataset

Our data set, *Bar Crawl: Detecting Heavy Drinking* (found here), captures UNIX timestamped accelerometer data from 19 study participants' phones with the hopes of using that data to determine whether the individual is drunk by way of capturing staggered steps (fluctuating values in x, y, and/or z) or falls (sharp variation in one or two axes). Each participant also has their real-time TAC and body temperature recorded in supplementary tables, the prior of which can be used for model testing purposes by comparing what a machine learning algorithm predicts to their real TAC level. TAC is recorded in supplementary tables by participant and is UNIX timestamped. Additionally, in order to control for variance in device readouts, the OS of each participants' phone is kept in an additional supplementary table. In sum, there are 14,057,567 accelerometer readings, 715 TAC readings, and 13 usable participants.

## Research Question

Our core research conceit is as follows: can we predict heavy drinking episodes by analyzing mobile phone accelerometer data? In practice, this sort of model would be invaluable to those who are prone to heavy drinking, allowing their devices to potentially warn them of their TAC without needing to draw blood or perform a breathalyzer—it would be provided dynamically and automatically as they progress through the night. Beyond drinking contexts, the answer to this question could be extrapolated onto elderly populations particularly at risk of falling and suffering a related injury. To answer this overarching question, several supplementary questions must be addressed first: How does TAC data correlate with the three-axis accelerometer time series data? What is the relationship between intoxication levels and accelerometer measurements? These questions will be answered naturally though the means described below.

# Multivariate Methods

## Clustering

We chose clustering as one of our multivariate methods as it would help us better identify underlying behavior types that may be indicative of heavy drinking episodes. By applying clustering, we can classify different movement patterns within the accelerometer data by grouping data points with similar movement patterns. We can then use this to explore patterns within the accelerometer data that correlate to TAC levels, giving us an effective way to differentiate behaviors related to intoxicated vs. non-intoxicated states.

## Multivariate normal modeling

Adjacent to clustering, a multivariate Normal model would allow us to assess large deviations from "normal" movement. This would be done via probability density functions and related contour plots. These methods would also allow us to establish a set of baseline behaviors and/or distributions between sober and intoxicated individuals.

# Initial observations about the data

The data seems ripe for a predictive machine learning algorithm. The supplementary TAC tables can be used as labels in the instance of a train_test_split()-using algorithm like multivariate logistic regression. That said, we worry about the sheer number of tables and how many of them will be used in the final proposal; each participant has a table unique to them capturing the TAC, the accelerometer dataset is 14 million rows long and contains every participants' data, and the additional dataset which labels each participant as an Android or iPhone user. We will likely either cut the Android data (being only two participants) or run two parallel models that analyze each phone type individually.

The data is also long—those aforementioned 14 million rows—meaning that processing time is likely going to be a concern. Any remotely robust machine learning model will take several hours to parse through that data, potentially days, meaning that the use of an external processing cluster would be invaluable. We can choose to reduce the number of rows ran at once by parsing through the data by participant, dropping the number of rows handled at once from 14 million to a little over 1 million, but that draws questions of whether the final model will be overfit on that single participant. While it is difficult to say which approach we will use presently, we will start with trying to parse the larger dataset, only resorting to parsing it participant-wise if the computation time is too large.

# Expected Contributions

**Aidan** will be responsible for EDA, data cleaning, the correlation matrix, any low-level predictive models like logistic regression, and general code hygiene.

**Alyssa** is our Aesthetics Lead. She will be responsible for ensuring visualizations are up to par both technically and aesthetically and the clustering multivariate method.

**Audrey** is analyzing our data to ensure that at least three of our columns follow a multivariate normal distribution.

**Connar** is doing multivariate normal model, generating contour plots and probability density functions, and any high-level predictive models if appropriate.