Action set

Action set in this project is related to the type of sensors we use to collected data. This is because we can use the collected data to set the relevant payoffs. For this project, we collect sleep data, step data, location data, smartphone data and computer screen usage data. Therefore, the action set could be increasing sleep, reducing the usage of smartphone and computer, strengthening excersize.

Raw data

We utilized a dataset of the first-year undergraduate students who were recruited by the for a health and well-being research study. These students come from an American University in the state of Pennsylvania. A total of 188 undergraduate students are enrolled in the data collection process, 160 (61% female, 57% Asian, 34% white, 9% Hispanic, and 5% black) of whom completed all the data collection tasks. The entire process of data collection and processing lasted an entire semester (approximately 16 weeks). Moreover, we guarantee that the privacy of each participant is fully protected. This data collection process adopts the anonymous collection method. We have assigned an identity document (ID) to each participant, and the ID will not be associated with the participant's personal information. Because the data type contains location information, the location code is used instead of the specific name of the location. Data collection methods include smartphones, smart wearable devices, and questionnaires. Smartphones are Android or Apple phones that belong to each participant's daily use. We ask participants to download the award framework app, which features hardware, software, And human-based data from smartphones and transforms it into information that you can understand. Call, message, and phone usage are event-based data, while Bluetooth, Wi-Fi, and location are sampled as time series. For the smart bracelet, after considering the ease of use, reliability, and price, we chose Fitbit flex 2 as the data collection tool. Like other Fitbit trackers, flex 2 can easily record people's most important health and fitness data, including steps, distances, calories burned, and sleep throughout the day. These data is sampled at different rate and we need to synchronize it before we use it. Questionnaires are web-based and used to measure the their mental health and well-being. There will be two same surveys during the entire data collection process, one is at the beginning of the semester and the other other is at the end of the semester. The data from AWARE will be automatically transmitted to our back-end server periodically via Wi-Fi, and the Fitbit application programming interface (API) will be used to retrieve wearable Fitbit data at the end of the study. Participants were required to keep their phones and Fitbit powered on as much as possible to reduce the proportion of missing data and ensure data continuity.

Feature extraction

Our data is a group of time series data output by multi-sensor. In order to apply machine learning algorithms to this time series data, we need the help of feature engineering. In addition, feature extraction can control the selection of essential and useful features by eliminating redundant features and noises in the data set so as to produce the best prediction output. At present, there are many feature engineering technologies for time series data, such as "tsfresh," an open-source Python library that can calculate all kinds of features. In this project, we use the feature extraction tool based

on the previous work: a generic and flexible feature extraction component (FEC). FEC, like most common feature extraction tools, uses some common statistical features, such as minimum, median, average, maximum, and standard deviation. The difference is that FEC defines specific characteristics for simultaneous interpreting data from different sensors. Our data comes from Bluetooth, Wi-Fi, communication, location, steps, sleep, and smartphone usage. For these sensors, FEC adds some complex behavioral features, such as day and night movement, sleep efficiency, and travel distance. FEC can calculate the characteristics of time series in different periods (from 5 minutes to several months). We get time segments from the whole dataset according to the specified period, such as noon, afternoon, and evening in the morning, the entire day, and weekdays and weekends. Then we extract features from these time segments. Finally, we obtain 77805 features from the combination of time series of different periods mentioned above.

State space

The state space consists of the rhythm statues of students. And we decide to use the rhythm to identify the health condtion of students. For example, normal student may have a 24 hour circadian rhythm, on the contrary some unhealthy students may have rhythms which may be less or more than 24 hours. We plan to use time windowing, and time windowing is one of the most frequently used processing methods for streams of data. A big dataset (events) is split into finite sets, or windows, based on specified criteria, such as time. In this way, the whole dataset could be divided into several time windows. Then, we will apply human rhythm detection algorithm on each time window.

Data generating process

We use the rhythm detection algorithm to identify rhythms according to history data. Based on the rhythms outputted by the rhythm dectection module, the rhythm in next time window could also be modeled. In fact, the prediction algorithm samples from a set of potential rhythms and feed back the one with maximal probability.

Rhythm detection algorithm

Periodogram

Periodogram provides a measure of the strength and regularity of the underlying rhythm. A periodogram is an estimate of the spectral density of a signal. Unlike cosinor to which the period should be known, the periodogram uses a Fourier Transform to convert a signal from the time domain to the frequency domain. A Fourier analysis is a method for expressing a function as a sum of periodic components, and for recovering the service from those components. The dominant frequency corresponds to the apparent periodic pattern.

Cosinor

Cosinor uses the least squares method to fit one or several cosine curves with or without polynomial terms to a single time series. Cosinor models have been used to characterize circadian rhythms and to compute relevant parameters with their confidence limits. Although unlike periodogram, the period

should be assumed known (or approximately known) for cosinor, the model outputs the significance of the period and it is proved that if $P \leq 0.05$, the assumed period actually exists. The outputs of Cosinor model include rhythmic parameters MESOR, amplitude, and acrophase (\textcolor{red}{add the rest of features here}. The model can be composed according to the equation:

$$x_i = M + Acos(t_i + \phi) + e_i$$

where M is the MESOR, A is the amplitude, t_i is the sampling time, ϕ is the acrophase and e is the error term. Besides, Cosinor outputs standard error for Mesor, amplitude, and acrophase respectively.

Rhythm Features

Period

The cycles of organisms in physical strength, emotion, and intelligence due to the influence of its own mechanism, nature, and environment are called biorhythm. And the time taken by a creature to complete one of this kind of cycle is called rhythm period. For example, circadian rhythm shows a continuous and stable periodicity of about 24 hours.

MESOR

Mesor is the midline of rhythm data calculated by the Cosinor model. When the sampling interval is equal, the MESOR is equal to the mean value of all rhythm data points.

Amplitude

Amplitude refers to the maximum value that fluctuating rhythm data can reach. The amplitude of a symmetrical wave is half of its range of up and down oscillation.

Acrophase

The log relative to reference time when rhythm data fluctuates to the apex is called orthophase.

Prior beliefs

People always hope to have a normal circadian rhythm. The normal circadian rhythm will contribute to efficient work and better life.

Payoffs

Since our sensors could collect data related to our action, we can get a scientific estimated payoff for a certain action through the analysis our collected dataset. For example, we know the sleep sensor data in a time window and can calculate the length of the rhythm for this window. Then, we also

know the change of the sleep datain the next window. Thus, the payoff of the change of sleep could be determined by whether the rhythm length changes and how much the the rhythm length changes.

Utility function or loss function

$U(a, X)$, where a represents the action and X represents the state. Here, the action is from the action set mentioned in the action set part and state is the rhythm status. We need to maximize the expection the utility function given the probability distribution of the next time window rhythms: $MaxE[U(a, X)|\theta]$.

Rule of decision makers

Generally speaking, peopele prefer to select the action which maximize benefits easily and conveniently. So, we can provide an order list of the potential actions according to accessibility and convenience.