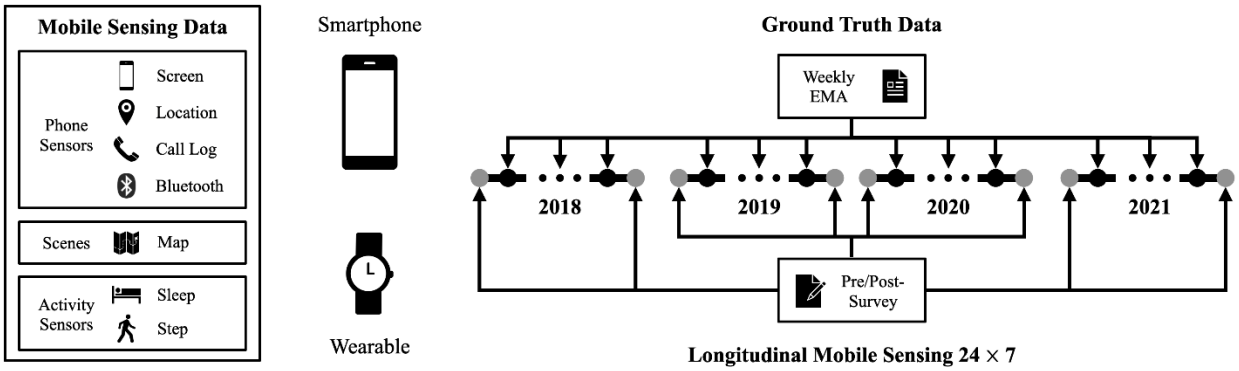
040613702 Machine Learning ปีการศึกษา 2568
คณะวิทยาศาสตร์ประยุกต์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Data for Course project

Data Collection

The overall data collection procedure is shown in the following figure:



The owner of dataset developed a mobile app using the AWARE framework that continuously collects location, phone usage (screen status), Bluetooth scans, and call logs. The app is compatible with both the iOS and Android platforms. Participants installed the app on smartphones and left it running in the background. In addition, we provided Fitbits to collect their physical activities and sleep behaviors. The mobile app and wearable passively collected sensor data 24$\times$7 during the study. The average study length is 78 days per person per year among the four datasets.

Meanwhile, surveys are delivered to participants at the start/end and during the study. These surveys cover a wide range of life experience of participants, including personality, physical well-being, mental well-being, social justice, and substance usage. Please check this website https://the-globem.github.io/datasets/overview for more information.

Target variable ::
['f_slp:fitbit_sleep_summary_rapids_avgefficiencymain:allday','f_slp:fitbit_sleep_summary_rapids_avgefficiencymain_norm:allday','f_slp:fitbit_sleep_summary_rapids_sumdurationasleepmain_norm:allday','f_slp:fitbit_sleep_summary_rapids_sumdurationasleepmain:allday']

Feature Data

The **FeatureData** folder contains seven files, all indexed by **pid** and **date**.

- rapids.csv: The complete feature file that contains all features.
- location.csv: The feature file that contains all location features.
- screen.csv: The feature file that contains all phone usage features.

- call.csv: The feature file that contains all call features.
- bluetooth.csv: The feature file that contains all Bluetooth features.
- steps.csv: The feature file that contains all physical activity features.
- sleep.csv: The feature file that contains all sleep features.
- wifi.csv: The feature file that contains all WiFi features. Note that this feature type is not used by any existing algorithms and often has a high data missing rate.

I already combine all seven file into one file (Sleep_data__nonorm_ML2568 and Sleep_data_norm_ML2568). Nonorm contain a raw value and norm contain normalized values.

Please note that all features are extracted with multiple **time_segment**s

- morning (6 am - 12 pm, calculated daily)
- afternoon (12 pm - 6 pm, calculated daily)
- evening (6 pm - 12 am, calculated daily)
- night (12 am - 6 am, calculated daily)
- allday (24 hrs from 12 am to 11:59 pm, calculated daily)
- 7-day history (calculated daily)
- 14-day history (calculated daily)
- weekdays (calculated once per week on Friday)

- weekend (calculated once per week on Sunday)

For all features with numeric values, we also provide two more versions:

- normalized: subtracted by each participant's median and divided by the 5-95 quantile range
- discretized: low/medium/high split by 33/66 quantile of each participant's feature value

Naming Format

All features follow a consistent naming format:

[feature_type]:[feature_name][version]:[time_segment]

- **feature_type**: It corresponds to the six data types.
    - location - f_loc
    - screen - f_screen
    - call - f_call
    - bluetooth - f_blue
    - steps - f_steps
    - sleep - f_slp.
- **feature_name**: The name of the feature provided by RAPIDS, i.e., the second column of the following figure, plus some additional information. A typical format is [SensorType]_[CodeProvider]_[featurename]. Please refer to RAPIDS's naming format [9] for more details.

- **version**: It has three versions:
  - 1) nothing, just empty "";
  - 2) normalized, _norm;
  - 3) discretized, _dis.
- **time_segment**: It corresponds to the specific time segment.
  - morning - morning
  - afternoon - afternoon
  - evening - evening
  - night - night
  - allday - allday
  - 7-day history - 7dhist
  - 14-day history - 14dhist
  - weekday - weekday
  - weekend - weekend

A participant's "sumdurationunlock" normalized feature in mornings is "f_loc:phone_screen_rapids_sumdurationunlock_norm:morning".

Please find the following tables about feature details in our datasets.

Location Details

Feature Name	Unit	Description
hometime	minutes	Time at home. Time spent at home in minutes. Home is the most visited significant location between 8 pm and 8 am, including any pauses within a 200-meter radius.
disttravelled	meters	Total distance traveled over a day (flights).
rog	meters	The Radius of Gyration (rog) is a measure in meters of the area covered by a person over a day. A centroid is calculated for all the places (pauses) visited during a day, and a weighted distance between all the places and that centroid is computed. The weights are proportional to the time spent in each place.
maxdiam	meters	The maximum diameter is the largest distance between any two pauses.
maxhomedist	meters	The maximum distance from home in meters.

siglocsvisited	locations	The number of significant locations visited during the day. Significant locations are computed using k-means clustering over pauses found in the whole monitoring period. The number of clusters is found iterating k from 1 to 200 stopping until the centroids of two significant locations are within 400 meters of one another.
avgflightlen	meters	Mean length of all flights.
stdflightlen	meters	Standard deviation of the length of all flights.
avgflightdur	seconds	Mean duration of all flights.
stdflightdur	seconds	The standard deviation of the duration of all flights.
probpause	-	The fraction of a day spent in a pause (as opposed to a flight).
siglocentropy	nats	Shannon’s entropy measurement is based on the proportion of time

		spent at each significant location visited during a day.
circdnrtn	-	A continuous metric quantifying a person's circadian routine that can take any value between 0 and 1, where 0 represents a daily routine completely different from any other sensed days and 1 a routine the same as every other sensed day.
wkenddayrtn	-	Same as circdnrtn but computed separately for weekends and weekdays.
locationvariance	meters2	The sum of the variances of the latitude and longitude columns.
loglocationvariance	-	Log of the sum of the variances of the latitude and longitude columns.
totaldistance	meters	Total distance traveled in a time segment using the haversine formula.
avgspeed	km/hr	Average speed in a time segment considering only the instances labeled as Moving. This feature is

		0 when the participant is stationary during a time segment.
varspeed	km/hr	Speed variance in a time segment considering only the instances labeled as Moving. This feature is 0 when the participant is stationary during a time segment.
numberofsignificantplaces	places	Number of significant locations visited. It is calculated using the DBSCAN/OPTICS clustering algorithm which takes in EPS and MIN_SAMPLES as parameters to identify clusters. Each cluster is a significant place.
numberlocationtransitions	transitions	Number of movements between any two clusters in a time segment.
radiusgyration	meters	Quantifies the area covered by a participant.
timeattop1location	minutes	Time spent at the most significant location.
timeattop2location	minutes	Time spent at the 2nd most significant location.

timeattop3location	minutes	Time spent at the 3rd most significant location.
movingtostaticratio	-	Ratio between stationary time and total location sensed time. A lat/long coordinate pair is labeled as stationary if its speed (distance/time) to the next coordinate pair is less than 1km/hr. A higher value represents a more stationary routine.
outlierstimepercent	-	Ratio between the time spent in non-significant clusters divided by the time spent in all clusters (stationary time. Only stationary samples are clustered). A higher value represents more time spent in non-significant clusters.
maxlengthstayatclusters	minutes	Maximum time spent in a cluster (significant location).
minlengthstayatclusters	minutes	Minimum time spent in a cluster (significant location).
avglengthstayatclusters	minutes	Average time spent in a cluster (significant location).

stdlengthstayatclusters	minutes	Standard deviation of time spent in a cluster (significant location).
locationentropy	nats	Shannon Entropy computed over the row count of each cluster (significant location), it is higher the more rows belong to a cluster (i.e., the more time a participant spent at a significant location).
normalizedlocationentropy	nats	Shannon Entropy computed over the row count of each cluster (significant location) divided by the number of clusters; it is higher the more rows belong to a cluster (i.e., the more time a participant spent at a significant location).
timeathome	minutes	Time spent at home.
timeat[PLACE]	minutes	Time spent at [PLACE], which can be living, exercise, study, greens.

Phone Usage Details

Feature Name	**Unit**	**Description**

sumduration	minutes	Total duration of all unlock episodes.
maxduration	minutes	Longest duration of any unlock episode.
minduration	minutes	Shortest duration of any unlock episode.
avgduration	minutes	Average duration of all unlock episodes.
stdduration	minutes	Standard deviation duration of all unlock episodes.
countepisode	episodes	Number of all unlock episodes.
firstuseafter	minutes	Minutes until the first unlock episode.
sumduration[PLACE]	minutes	Total duration of all unlock episodes. [PLACE] can be living, exercise, study, greens. Same below.
maxduration[PLACE]	minutes	Longest duration of any unlock episode.

minduration[PLACE]	minutes	Shortest duration of any unlock episode.
avgduration[PLACE]	minutes	Average duration of all unlock episodes.
stdduration[PLACE]	minutes	Standard deviation duration of all unlock episodes.
countepisode[PLACE]	episodes	Number of all unlock episodes.
firstuseafter[PLACE]	minutes	Minutes until the first unlock episode.

Call Details

Feature Name	**Unit**	**Description**
count	calls	Number of calls of a particular call_type (incoming/outgoing) occurred during a particular time_segment.
distinctcontacts	contacts	Number of distinct contacts that are associated with a particular call_type for a particular time_segment.

meanduration	seconds	The mean duration of all calls of a particular call_type during a particular time_segment.
sumduration	seconds	The sum of the duration of all calls of a particular call_type during a particular time_segment.
minduration	seconds	The duration of the shortest call of a particular call_type during a particular time_segment.
maxduration	seconds	The duration of the longest call of a particular call_type during a particular time_segment.
stdduration	seconds	The standard deviation of the duration of all the calls of a particular call_type during a particular time_segment.
modeduration	seconds	The mode of the duration of all the calls of a particular call_type during a particular time_segment.
entropyduration	nats	The estimate of the Shannon entropy for the the duration of all the

		calls of a particular call_type during a particular time_segment.
timefirstcall	minutes	The time in minutes between 12:00am (midnight) and the first call of call_type.
timelastcall	minutes	The time in minutes between 12:00am (midnight) and the last call of call_type.
countmostfrequentcontact	calls	The number of calls of a particular call_type during a particular time_segment of the most frequent contact throughout the monitored period.

Bluetooth Details

Feature Name	**Unit**	**Description**
countscans	scans	Number of scans (rows) from the devices sensed during a time segment instance. The more scans a bluetooth device has the longer it remained within range

		of the participant's phone.
uniquedevices	devices	Number of unique bluetooth devices sensed during a time segment instance as identified by their hardware addresses.
meanscans	scans	Mean of the scans of every sensed device within each time segment instance.
stdscans	scans	Standard deviation of the scans of every sensed device within each time segment instance.
countscansmostfrequentdevicewithinsegments	scans	Number of scans of the most sensed device within each time segment instance.
countscansleastfrequentdevicewithinsegments	scans	Number of scans of the least sensed device

		within each time segment instance.
countscansmostfrequentdeviceacrosssegments	scans	Number of scans of the most sensed device across time segment instances of the same type.
countscansleastfrequentdeviceacrosssegments	scans	Number of scans of the least sensed device across time segment instances of the same type per device.
countscansmostfrequentdeviceacrossdataset	scans	Number of scans of the most sensed device across the entire dataset of every participant.
countscansleastfrequentdeviceacrossdataset	scans	Number of scans of the least sensed device across the entire dataset of every participant.

WiFi Details

Feature Name	Unit	Description
countscans	devices	Number of scanned WiFi access points connected during a time_segment, an access point can be detected multiple times over time and these appearances are counted separately.
uniquedevices	devices	Number of unique access point during a time_segment as identified by their hardware address.
countscansmostuniquedevice	scans	Number of scans of the most scanned access point during a time_segment across the whole monitoring period.

Physical Activity Details

Feature Name	Unit	Description
maxsumsteps	steps	The maximum daily step count during a time segment.
minsumsteps	steps	The minimum daily step count during a time segment.

avgsumsteps	steps	The average daily step count during a time segment.
mediansumsteps	steps	The median of daily step count during a time segment.
stdsumsteps	steps	The standard deviation of daily step count during a time segment.
sumsteps	steps	The total step count during a time segment.
maxsteps	steps	The maximum step count during a time segment.
minsteps	steps	The minimum step count during a time segment.
avgsteps	steps	The average step count during a time segment.
stdsteps	steps	The standard deviation of step count during a time segment.
countepisodesedentarybout	bouts	Number of sedentary bouts during a time segment.
sumdurationsedentarybout	minutes	Total duration of all sedentary bouts during a time segment.

maxdurationsedentarybout	minutes	The maximum duration of any sedentary bout during a time segment.
mindurationsedentarybout	minutes	The minimum duration of any sedentary bout during a time segment.
avgdurationsedentarybout	minutes	The average duration of sedentary bouts during a time segment.
stddurationsedentarybout	minutes	The standard deviation of the duration of sedentary bouts during a time segment.
countepisodeactivebout	bouts	Number of active bouts during a time segment.
sumdurationactivebout	minutes	Total duration of all active bouts during a time segment.
maxdurationactivebout	minutes	The maximum duration of any active bout during a time segment.
mindurationactivebout	minutes	The minimum duration of any active bout during a time segment.
avgdurationactivebout	minutes	The average duration of active bouts during a time segment.

stddurationactivebout	minutes	The standard deviation of the duration of active bouts during a time segment.

Sleep Details

Feature Name	Unit	Description
countepisode[LEVEL][TYPE]	episodes	Number of [LEVEL][TYPE] sleep episodes. [LEVEL] is one of awake and asleep and [TYPE] is one of main, nap, and all. Same below.
sumduration[LEVEL][TYPE]	minutes	Total duration of all [LEVEL][TYPE] sleep episodes.
maxduration[LEVEL][TYPE]	minutes	Longest duration of any [LEVEL][TYPE] sleep episode.
minduration[LEVEL][TYPE]	minutes	Shortest duration of any [LEVEL][TYPE] sleep episode.
avgduration[LEVEL][TYPE]	minutes	Average duration of all [LEVEL][TYPE] sleep episodes.
medianduration[LEVEL][TYPE]	minutes	Median duration of all [LEVEL][TYPE] sleep episodes.

stdduration[LEVEL][TYPE]	minutes	Standard deviation duration of all [LEVEL][TYPE] sleep episodes.
firstwaketimeTYPE	minutes	First wake time for a certain sleep type during a time segment. Wake time is number of minutes after midnight of a sleep episode's end time.
lastwaketimeTYPE	minutes	Last wake time for a certain sleep type during a time segment. Wake time is number of minutes after midnight of a sleep episode's end time.
firstbedtimeTYPE	minutes	First bedtime for a certain sleep type during a time segment. Bedtime is number of minutes after midnight of a sleep episode's start time.
lastbedtimeTYPE	minutes	Last bedtime for a certain sleep type during a time segment. Bedtime is number of minutes after midnight of a sleep episode's start time.

countepisodeTYPE	episodes	Number of sleep episodes for a certain sleep type during a time segment.
avgefficiencyTYPE	scores	Average sleep efficiency for a certain sleep type during a time segment.
sumdurationafterwakeupTYPE	minutes	Total duration the user stayed in bed after waking up for a certain sleep type during a time segment.
sumdurationasleepTYPE	minutes	Total sleep duration for a certain sleep type during a time segment.
sumdurationawakeTYPE	minutes	Total duration the user stayed awake but still in bed for a certain sleep type during a time segment.
sumdurationtofallasleepTYPE	minutes	Total duration the user spent to fall asleep for a certain sleep type during a time segment.
sumdurationinbedTYPE	minutes	Total duration the user stayed in bed (sumdurationtofallasleep + sumdurationawake + sumdurationasleep + sumdurationafterwakeup) for a

		certain sleep type during a time segment.
avgdurationafterwakeupTYPE	minutes	Average duration the user stayed in bed after waking up for a certain sleep type during a time segment.
avgdurationasleepTYPE	minutes	Average sleep duration for a certain sleep type during a time segment.
avgdurationawakeTYPE	minutes	Average duration the user stayed awake but still in bed for a certain sleep type during a time segment.
avgdurationtofallasleepTYPE	minutes	Average duration the user spent to fall asleep for a certain sleep type during a time segment.
avgdurationinbedTYPE	minutes	Average duration the user stayed in bed (sumdurationtofallasleep + sumdurationawake + sumdurationasleep + sumdurationafterwakeup) for a certain sleep type during a time segment.