2019

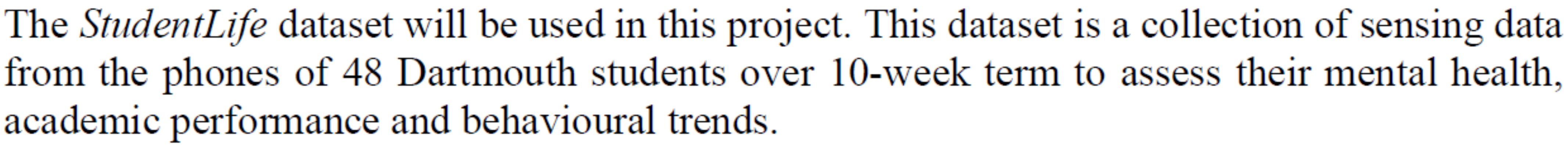
Comp9417

Project

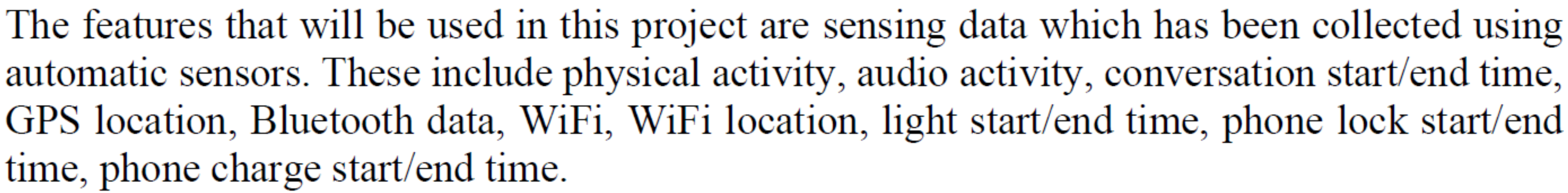
Comp9417
Final

Comp9417

Dataset Analysis

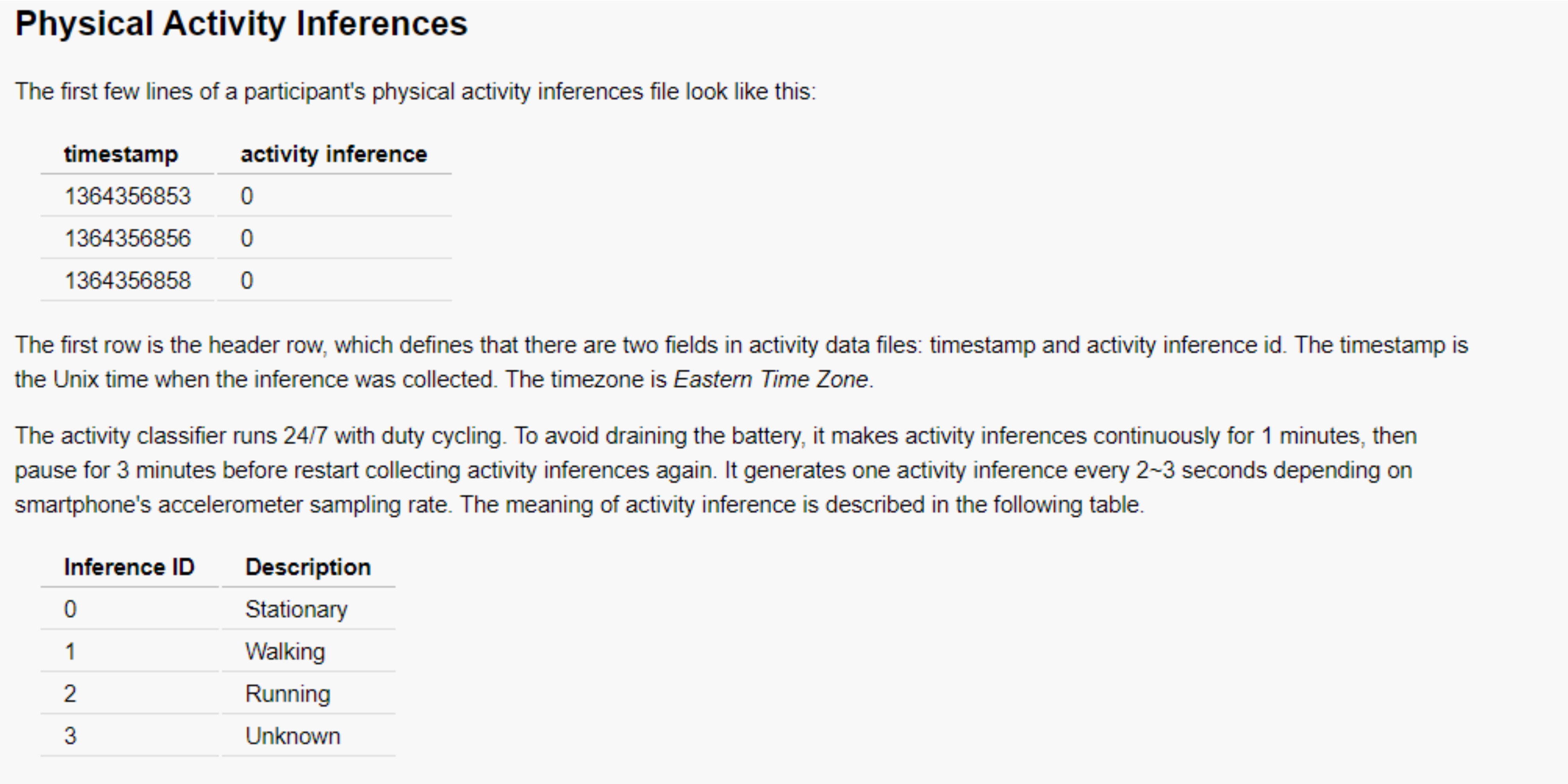


The *StudentLife* dataset will be used in this project. This dataset is a collection of sensing data from the phones of 48 Dartmouth students over 10-week term to assess their mental health, academic performance and behavioural trends.



The features that will be used in this project are sensing data which has been collected using automatic sensors. These include physical activity, audio activity, conversation start/end time, GPS location, Bluetooth data, WiFi, WiFi location, light start/end time, phone lock start/end time, phone charge start/end time.

Comp9417
Dataset Analysis



Physical Activity Inferences

The first few lines of a participant's physical activity inferences file look like this:

timestamp	activity inference
1364356853	0
1364356856	0
1364356858	0

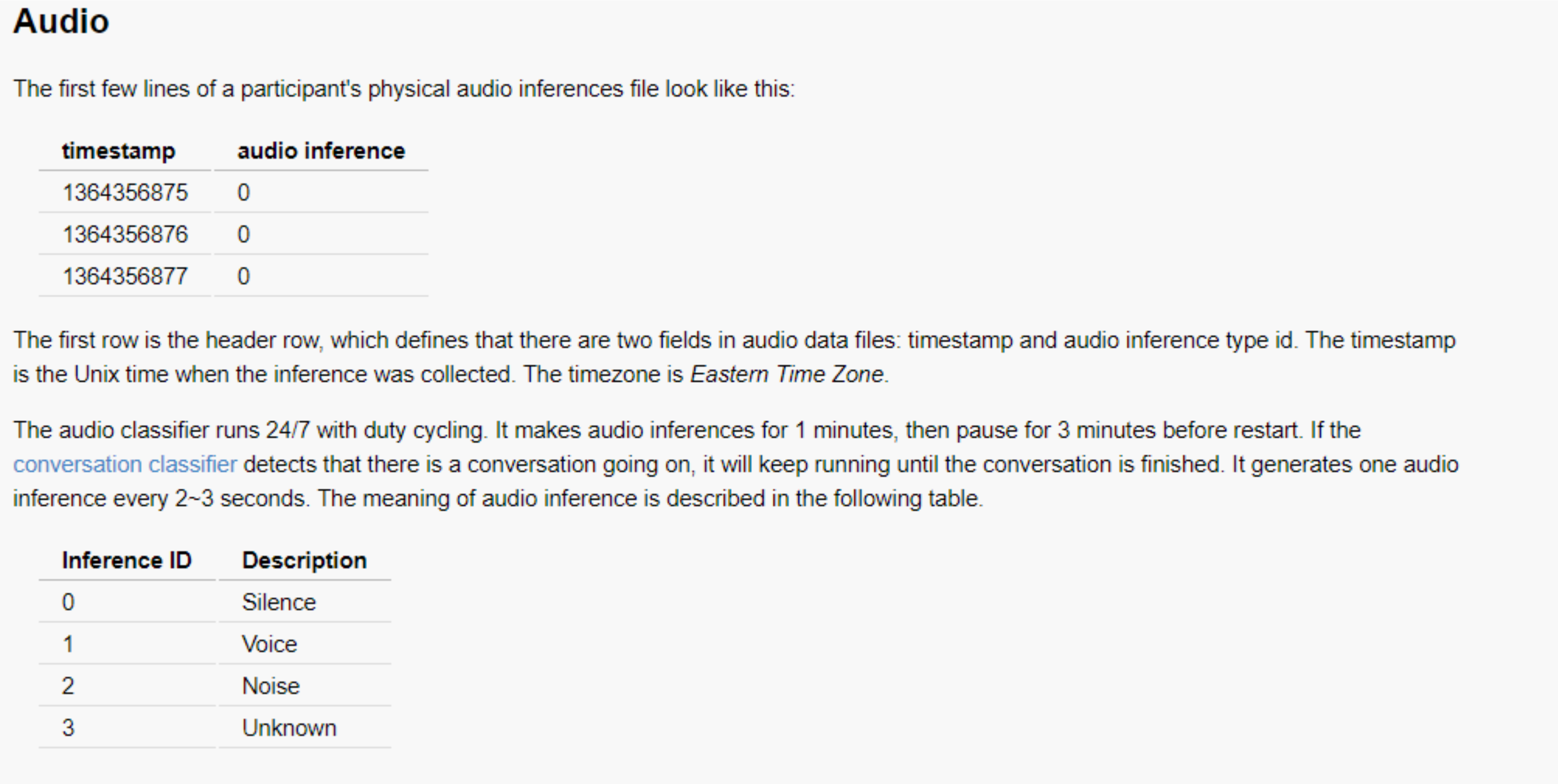
The first row is the header row, which defines that there are two fields in activity data files: timestamp and activity inference id. The timestamp is the Unix time when the inference was collected. The timezone is *Eastern Time Zone*.

The activity classifier runs 24/7 with duty cycling. To avoid draining the battery, it makes activity inferences continuously for 1 minutes, then pause for 3 minutes before restart collecting activity inferences again. It generates one activity inference every 2~3 seconds depending on smartphone's accelerometer sampling rate. The meaning of activity inference is described in the following table.

Inference ID	Description
0	Stationary
1	Walking
2	Running
3	Unknown

Comp9417

Dataset Analysis



Audio

The first few lines of a participant's physical audio inferences file look like this:

timestamp	audio inference
1364356875	0
1364356876	0
1364356877	0

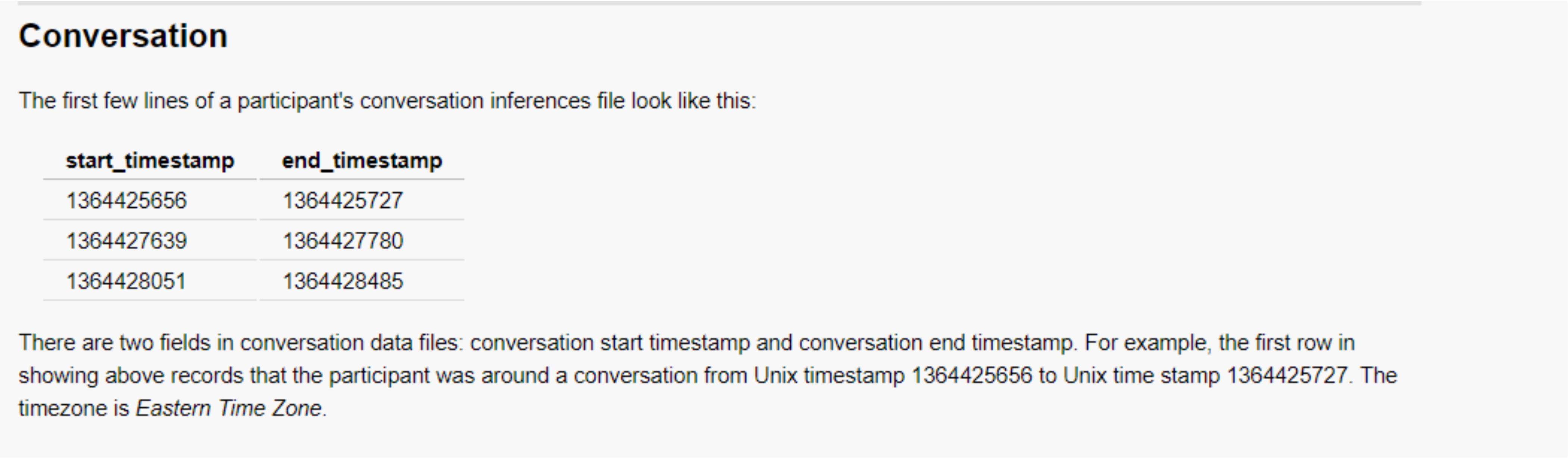
The first row is the header row, which defines that there are two fields in audio data files: timestamp and audio inference type id. The timestamp is the Unix time when the inference was collected. The timezone is *Eastern Time Zone*.

The audio classifier runs 24/7 with duty cycling. It makes audio inferences for 1 minutes, then pause for 3 minutes before restart. If the conversation classifier detects that there is a conversation going on, it will keep running until the conversation is finished. It generates one audio inference every 2~3 seconds. The meaning of audio inference is described in the following table.

Inference ID	Description
0	Silence
1	Voice
2	Noise
3	Unknown

Comp9417

Dataset Analysis



Conversation

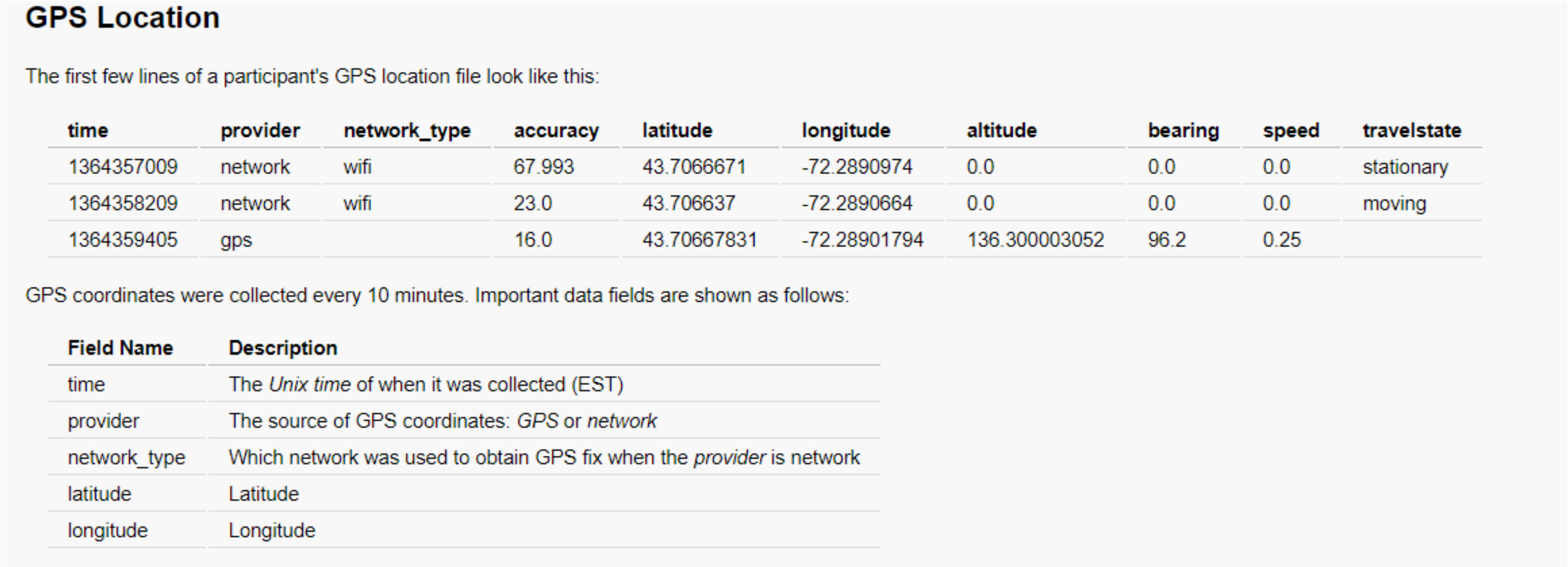
The first few lines of a participant's conversation inferences file look like this:

start_timestamp	end_timestamp
1364425656	1364425727
1364427639	1364427780
1364428051	1364428485

There are two fields in conversation data files: conversation start timestamp and conversation end timestamp. For example, the first row in showing above records that the participant was around a conversation from Unix timestamp 1364425656 to Unix time stamp 1364425727. The timezone is *Eastern Time Zone*.

Comp9417

Dataset Analysis



GPS Location

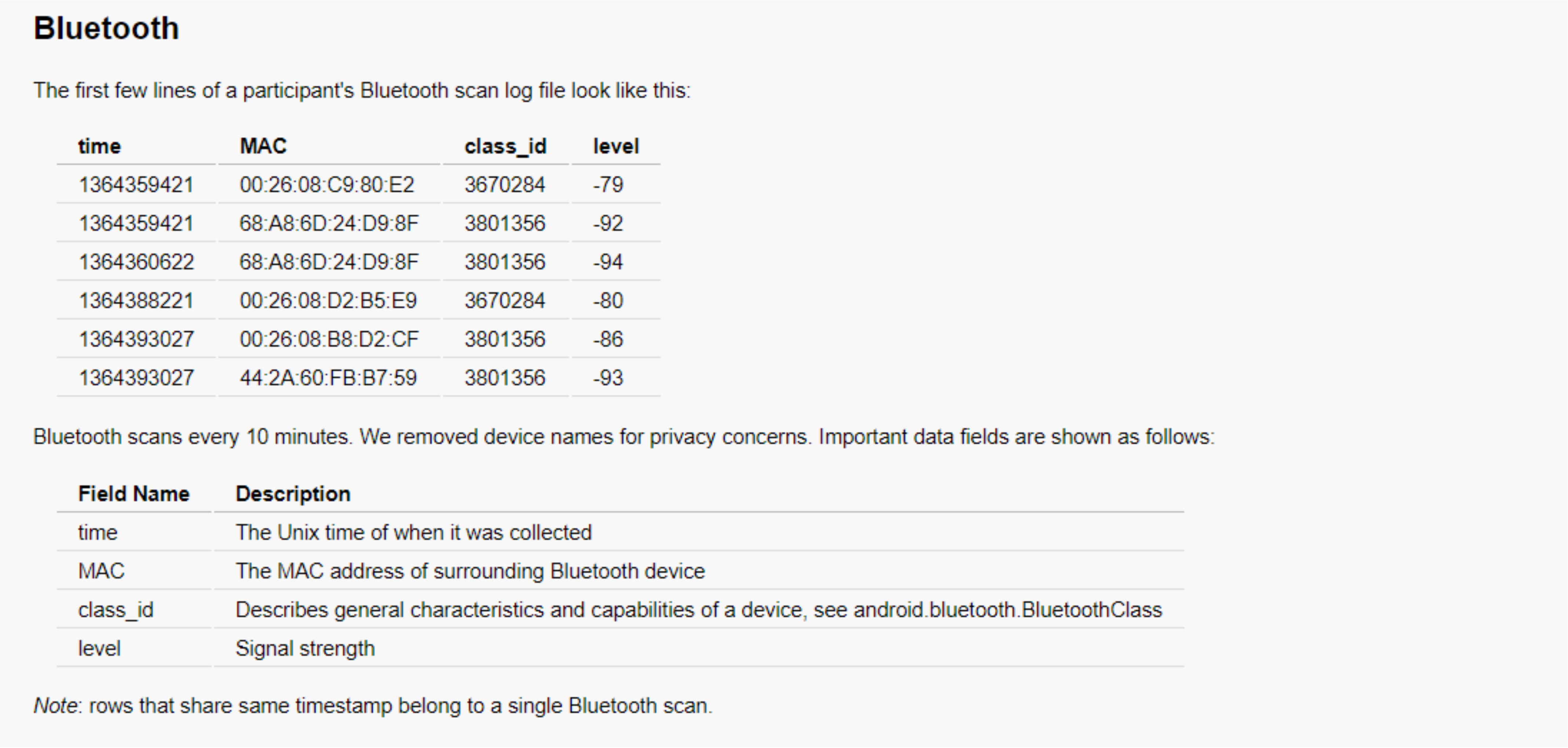
The first few lines of a participant's GPS location file look like this:

time	provider	network_type	accuracy	latitude	longitude	altitude	bearing	speed	travelstate
1364357009	network	wifi	67.993	43.7066671	-72.2890974	0.0	0.0	0.0	stationary
1364358209	network	wifi	23.0	43.706637	-72.2890664	0.0	0.0	0.0	moving
1364359405	gps		16.0	43.70667831	-72.28901794	136.300003052	96.2	0.25	

GPS coordinates were collected every 10 minutes. Important data fields are shown as follows:

Field Name	Description
time	The *Unix time* of when it was collected (EST)
provider	The source of GPS coordinates: *GPS* or *network*
network_type	Which network was used to obtain GPS fix when the *provider* is network
latitude	Latitude
longitude	Longitude

Comp9417
Dataset Analysis



Bluetooth

The first few lines of a participant's Bluetooth scan log file look like this:

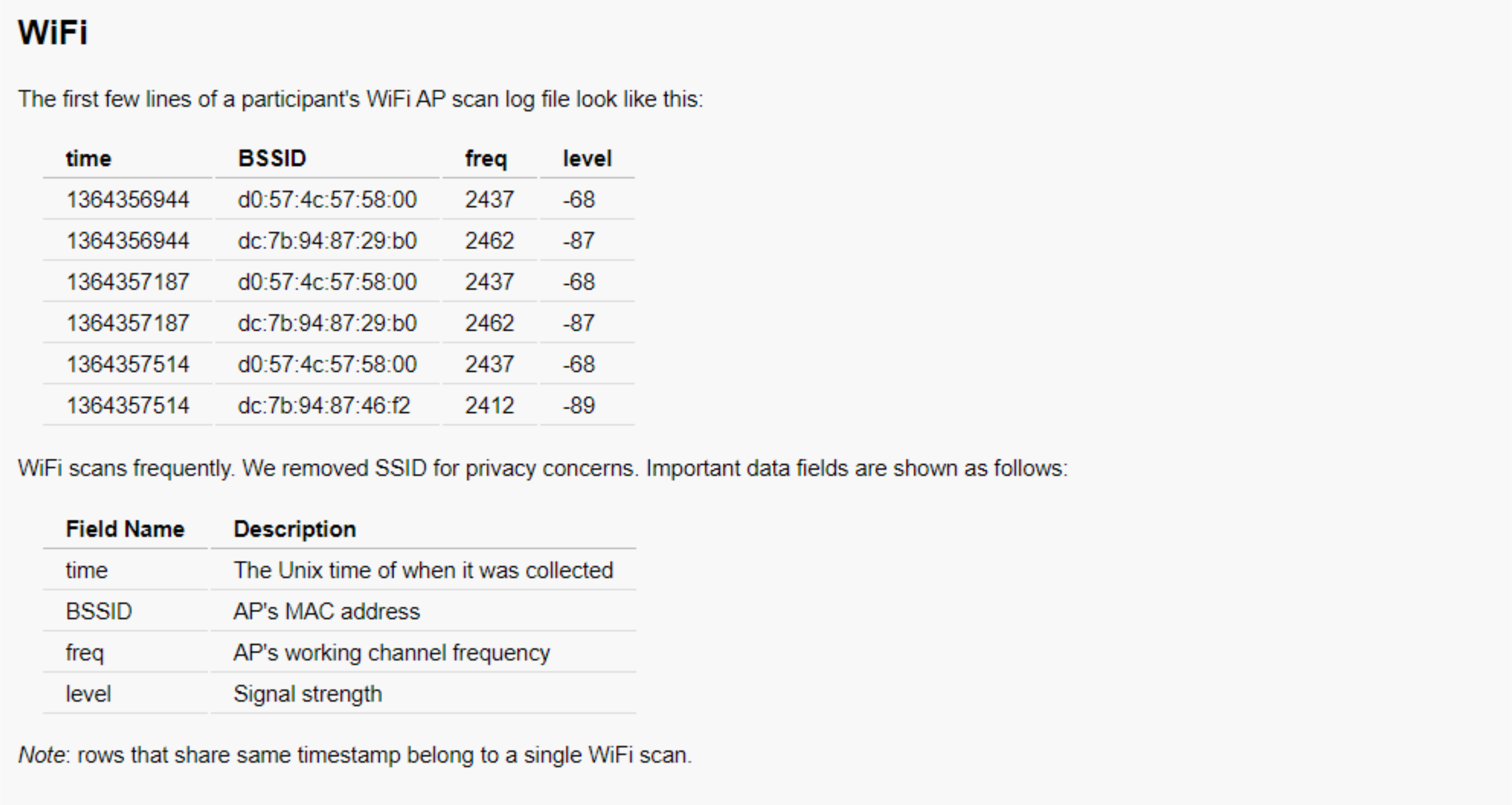
time	MAC	class_id	level
1364359421	00:26:08:C9:80:E2	3670284	-79
1364359421	68:A8:6D:24:D9:8F	3801356	-92
1364360622	68:A8:6D:24:D9:8F	3801356	-94
1364388221	00:26:08:D2:B5:E9	3670284	-80
1364393027	00:26:08:B8:D2:CF	3801356	-86
1364393027	44:2A:60:FB:B7:59	3801356	-93

Bluetooth scans every 10 minutes. We removed device names for privacy concerns. Important data fields are shown as follows:

Field Name	Description
time	The Unix time of when it was collected
MAC	The MAC address of surrounding Bluetooth device
class_id	Describes general characteristics and capabilities of a device, see android.bluetooth.BluetoothClass
level	Signal strength

Note: rows that share same timestamp belong to a single Bluetooth scan.

Comp9417
Dataset Analysis



WiFi

The first few lines of a participant's WiFi AP scan log file look like this:

time	BSSID	freq	level
1364356944	d0:57:4c:57:58:00	2437	-68
1364356944	dc:7b:94:87:29:b0	2462	-87
1364357187	d0:57:4c:57:58:00	2437	-68
1364357187	dc:7b:94:87:29:b0	2462	-87
1364357514	d0:57:4c:57:58:00	2437	-68
1364357514	dc:7b:94:87:46:f2	2412	-89

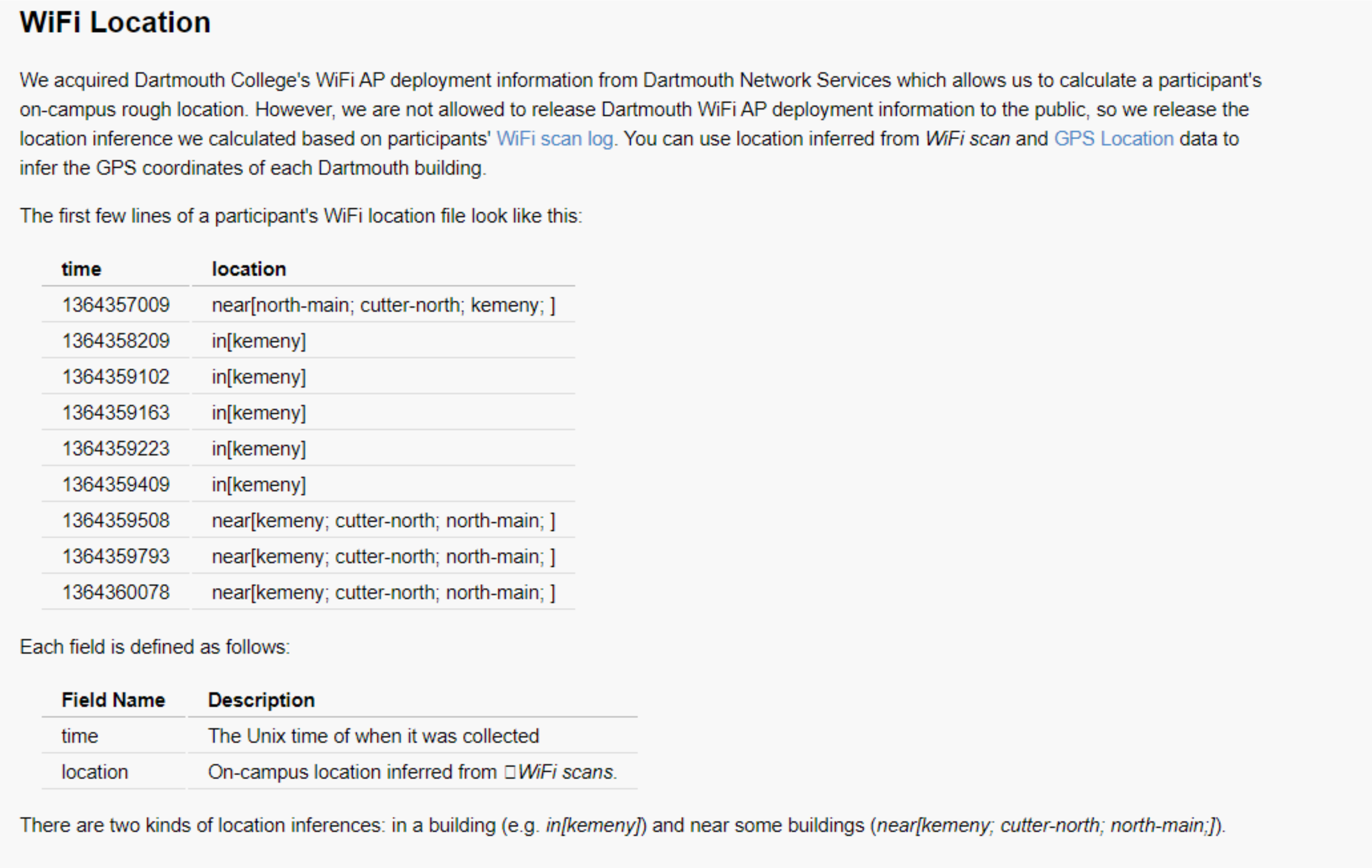
WiFi scans frequently. We removed SSID for privacy concerns. Important data fields are shown as follows:

Field Name	Description
time	The Unix time of when it was collected
BSSID	AP's MAC address
freq	AP's working channel frequency
level	Signal strength

Note: rows that share same timestamp belong to a single WiFi scan.

Comp9417

Dataset Analysis



WiFi Location

We acquired Dartmouth College's WiFi AP deployment information from Dartmouth Network Services which allows us to calculate a participant's on-campus rough location. However, we are not allowed to release Dartmouth WiFi AP deployment information to the public, so we release the location inference we calculated based on participants' WiFi scan log. You can use location inferred from *WiFi scan* and GPS Location data to infer the GPS coordinates of each Dartmouth building.

The first few lines of a participant's WiFi location file look like this:

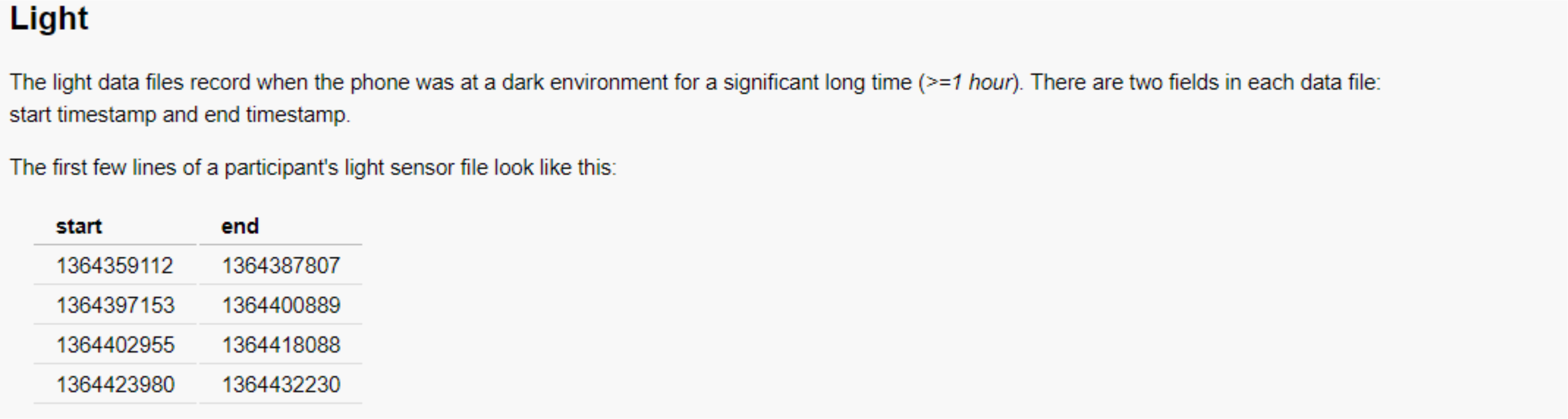
time	location
1364357009	near[north-main; cutter-north; kemeny;]
1364358209	in[kemeny]
1364359102	in[kemeny]
1364359163	in[kemeny]
1364359223	in[kemeny]
1364359409	in[kemeny]
1364359508	near[kemeny; cutter-north; north-main;]
1364359793	near[kemeny; cutter-north; north-main;]
1364360078	near[kemeny; cutter-north; north-main;]

Each field is defined as follows:

Field Name	Description
time	The Unix time of when it was collected
location	On-campus location inferred from □*WiFi scans*.

There are two kinds of location inferences: in a building (e.g. *in[kemeny]*) and near some buildings (*near[kemeny; cutter-north; north-main;]*).

Comp9417
Dataset Analysis



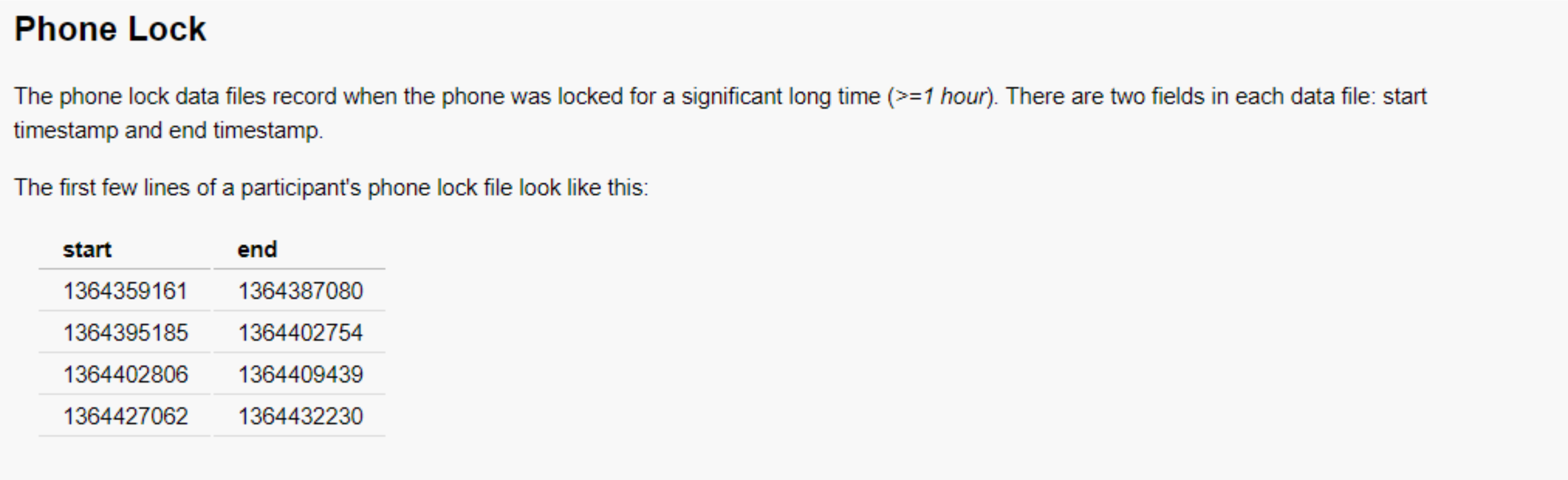
Light

The light data files record when the phone was at a dark environment for a significant long time (>=*1 hour*). There are two fields in each data file: start timestamp and end timestamp.

The first few lines of a participant's light sensor file look like this:

start	end
1364359112	1364387807
1364397153	1364400889
1364402955	1364418088
1364423980	1364432230

Comp9417
Dataset Analysis



Phone Lock

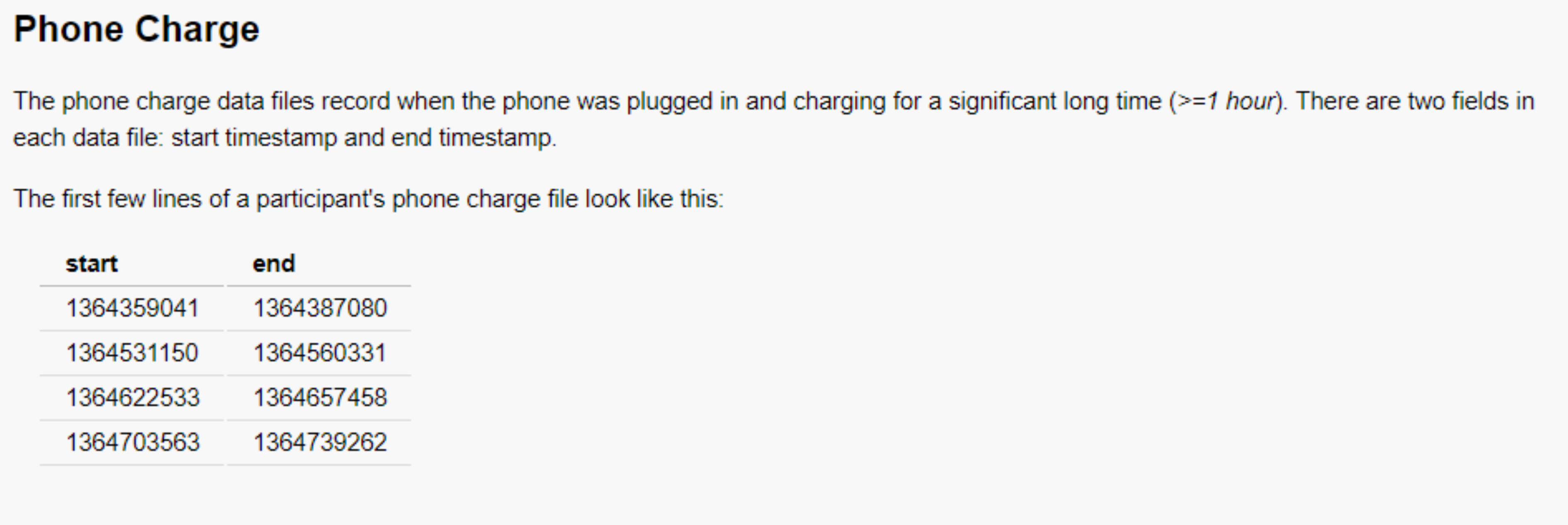
The phone lock data files record when the phone was locked for a significant long time (>=*1 hour*). There are two fields in each data file: start timestamp and end timestamp.

The first few lines of a participant's phone lock file look like this:

start	end
1364359161	1364387080
1364395185	1364402754
1364402806	1364409439
1364427062	1364432230

Comp9417

Dataset Analysis



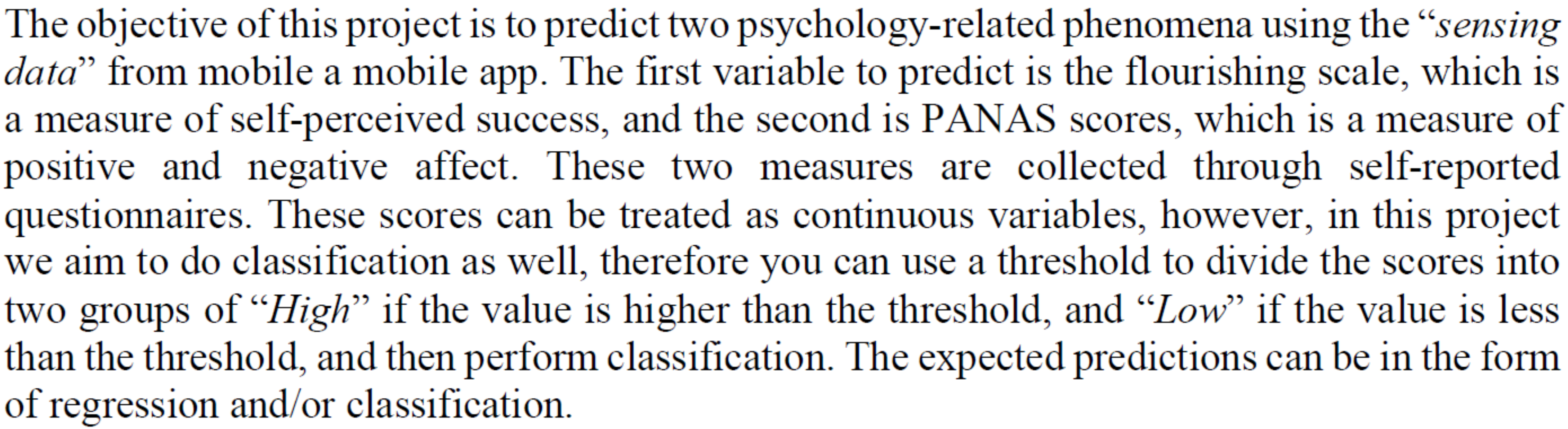
Phone Charge

The phone charge data files record when the phone was plugged in and charging for a significant long time (*>=1 hour*). There are two fields in each data file: start timestamp and end timestamp.

The first few lines of a participant's phone charge file look like this:

start	end
1364359041	1364387080
1364531150	1364560331
1364622533	1364657458
1364703563	1364739262

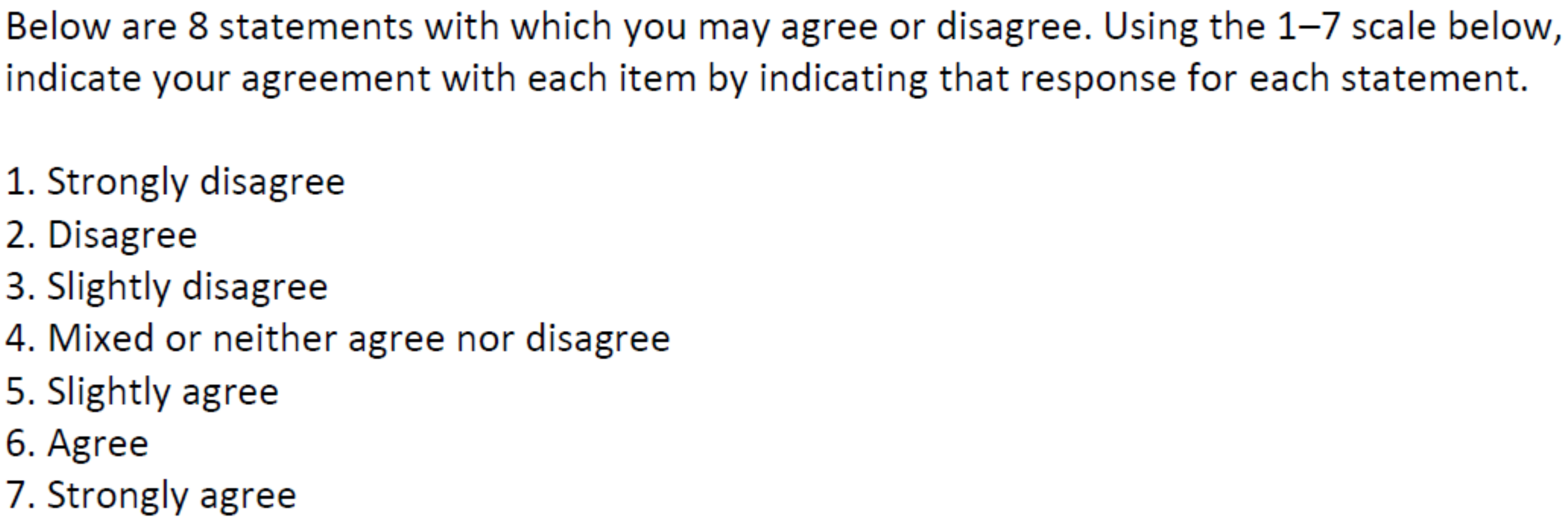
Comp9417
labels



The objective of this project is to predict two psychology-related phenomena using the “*sensing data*” from mobile a mobile app. The first variable to predict is the flourishing scale, which is a measure of self-perceived success, and the second is PANAS scores, which is a measure of positive and negative affect. These two measures are collected through self-reported questionnaires. These scores can be treated as continuous variables, however, in this project we aim to do classification as well, therefore you can use a threshold to divide the scores into two groups of “*High*” if the value is higher than the threshold, and “*Low*” if the value is less than the threshold, and then perform classification. The expected predictions can be in the form of regression and/or classification.

Comp9417

Flourishing scale



Below are 8 statements with which you may agree or disagree. Using the 1–7 scale below, indicate your agreement with each item by indicating that response for each statement.

1. Strongly disagree
2. Disagree
3. Slightly disagree
4. Mixed or neither agree nor disagree
5. Slightly agree
6. Agree
7. Strongly agree

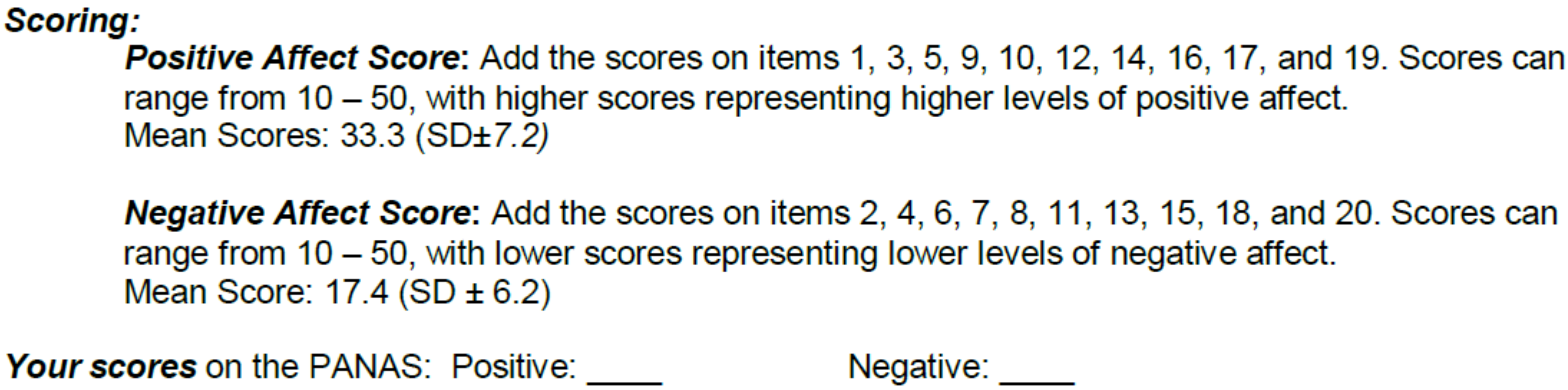
Comp9417
Flourishing scale

I lead a purposeful and meaningful life.
My social relationships are supportive and rewarding.
I am engaged and interested in my daily activities
I actively contribute to the happiness and well-being of others
I am competent and capable in the activities that are important to me
I am a good person and live a good life
I am optimistic about my future
People respect me

Scoring: Add the responses, varying from 1 to 7, for all eight items. The possible range of scores is from 8 (lowest possible) to 56 (highest PWB possible). A high score represents a person with many psychological resources and strengths.

Comp9417

Panas



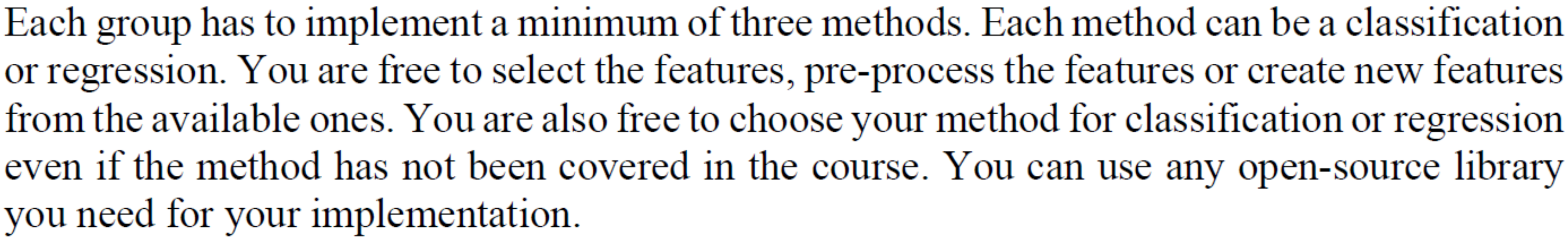
Scoring:

***Positive Affect Score*:** Add the scores on items 1, 3, 5, 9, 10, 12, 14, 16, 17, and 19. Scores can range from 10 – 50, with higher scores representing higher levels of positive affect.
Mean Scores: 33.3 (SD±*7.2)*

***Negative Affect Score*:** Add the scores on items 2, 4, 6, 7, 8, 11, 13, 15, 18, and 20. Scores can range from 10 – 50, with lower scores representing lower levels of negative affect.
Mean Score: 17.4 (SD ± 6.2)

Your scores on the PANAS: Positive: ____ Negative: ____

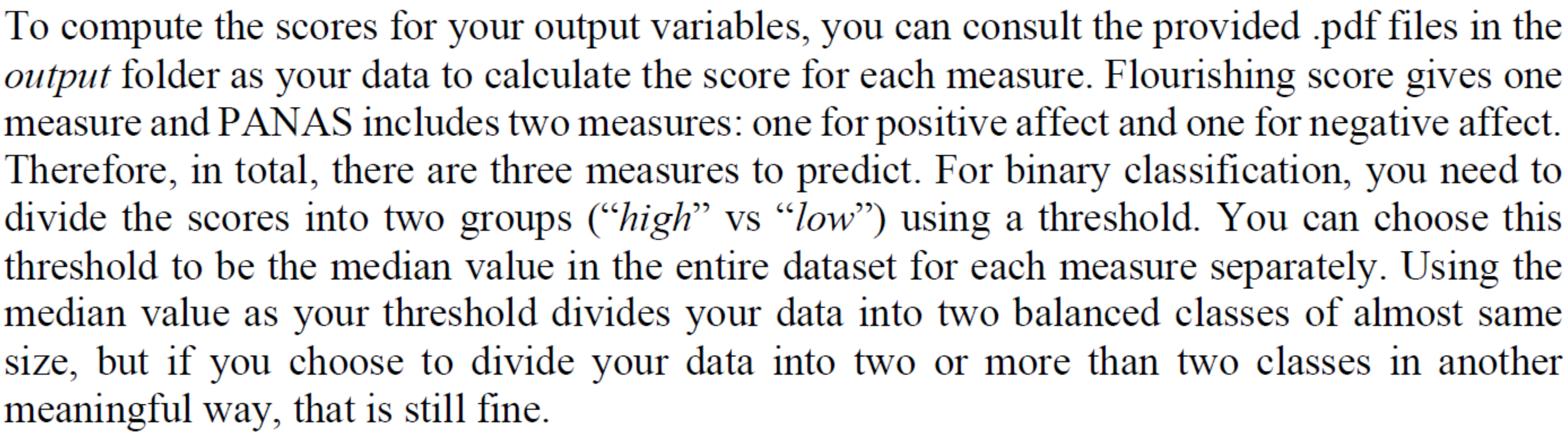
Comp9417
Methods



Each group has to implement a minimum of three methods. Each method can be a classification or regression. You are free to select the features, pre-process the features or create new features from the available ones. You are also free to choose your method for classification or regression even if the method has not been covered in the course. You can use any open-source library you need for your implementation.

Comp9417

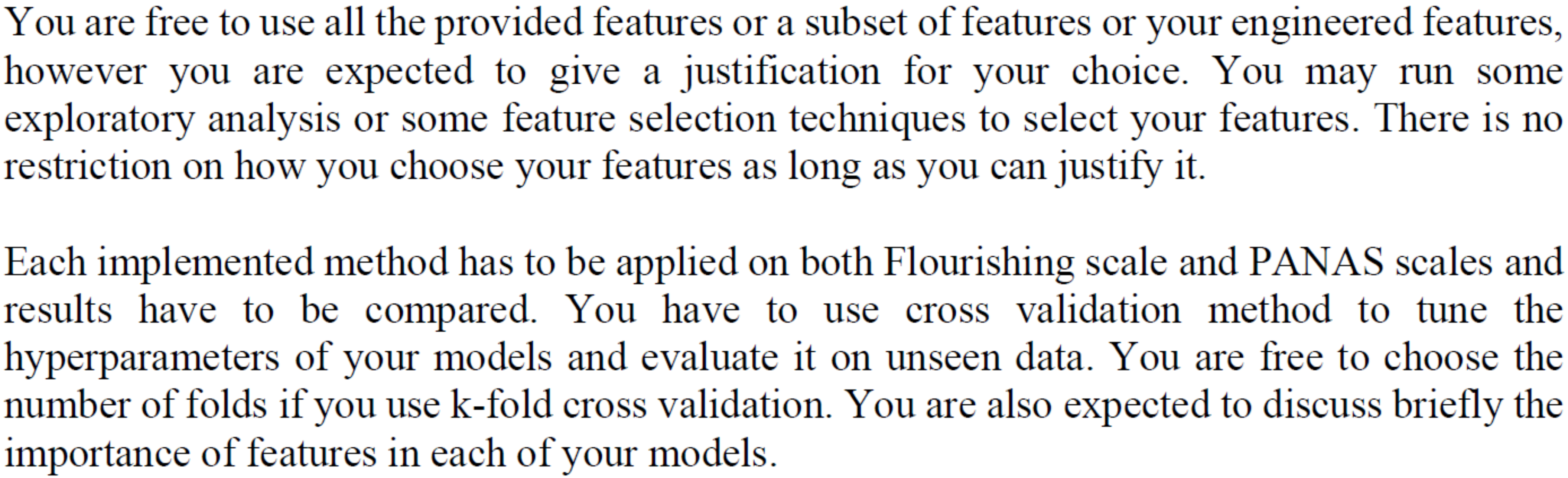
Methods



To compute the scores for your output variables, you can consult the provided .pdf files in the *output* folder as your data to calculate the score for each measure. Flourishing score gives one measure and PANAS includes two measures: one for positive affect and one for negative affect. Therefore, in total, there are three measures to predict. For binary classification, you need to divide the scores into two groups ("*high*" vs "*low*") using a threshold. You can choose this threshold to be the median value in the entire dataset for each measure separately. Using the median value as your threshold divides your data into two balanced classes of almost same size, but if you choose to divide your data into two or more than two classes in another meaningful way, that is still fine.

Comp9417

Methods

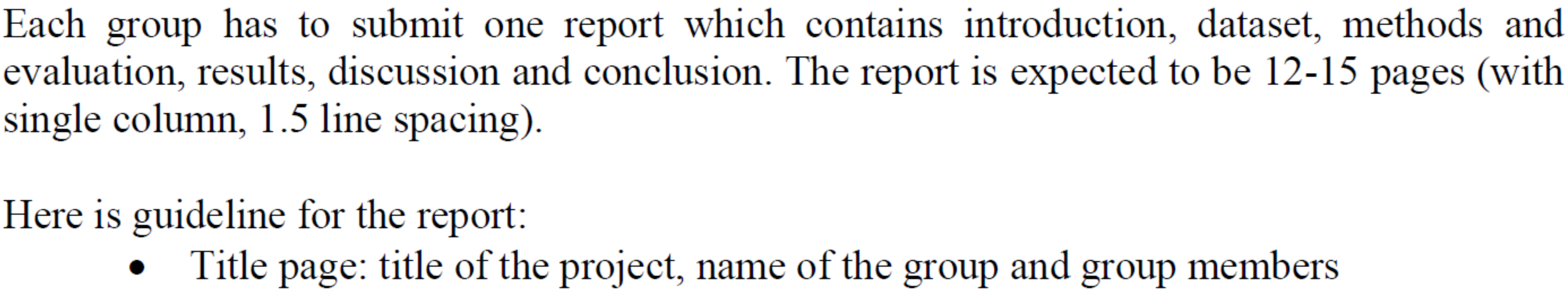


You are free to use all the provided features or a subset of features or your engineered features, however you are expected to give a justification for your choice. You may run some exploratory analysis or some feature selection techniques to select your features. There is no restriction on how you choose your features as long as you can justify it.

Each implemented method has to be applied on both Flourishing scale and PANAS scales and results have to be compared. You have to use cross validation method to tune the hyperparameters of your models and evaluate it on unseen data. You are free to choose the number of folds if you use k-fold cross validation. You are also expected to discuss briefly the importance of features in each of your models.

Comp9417

Report

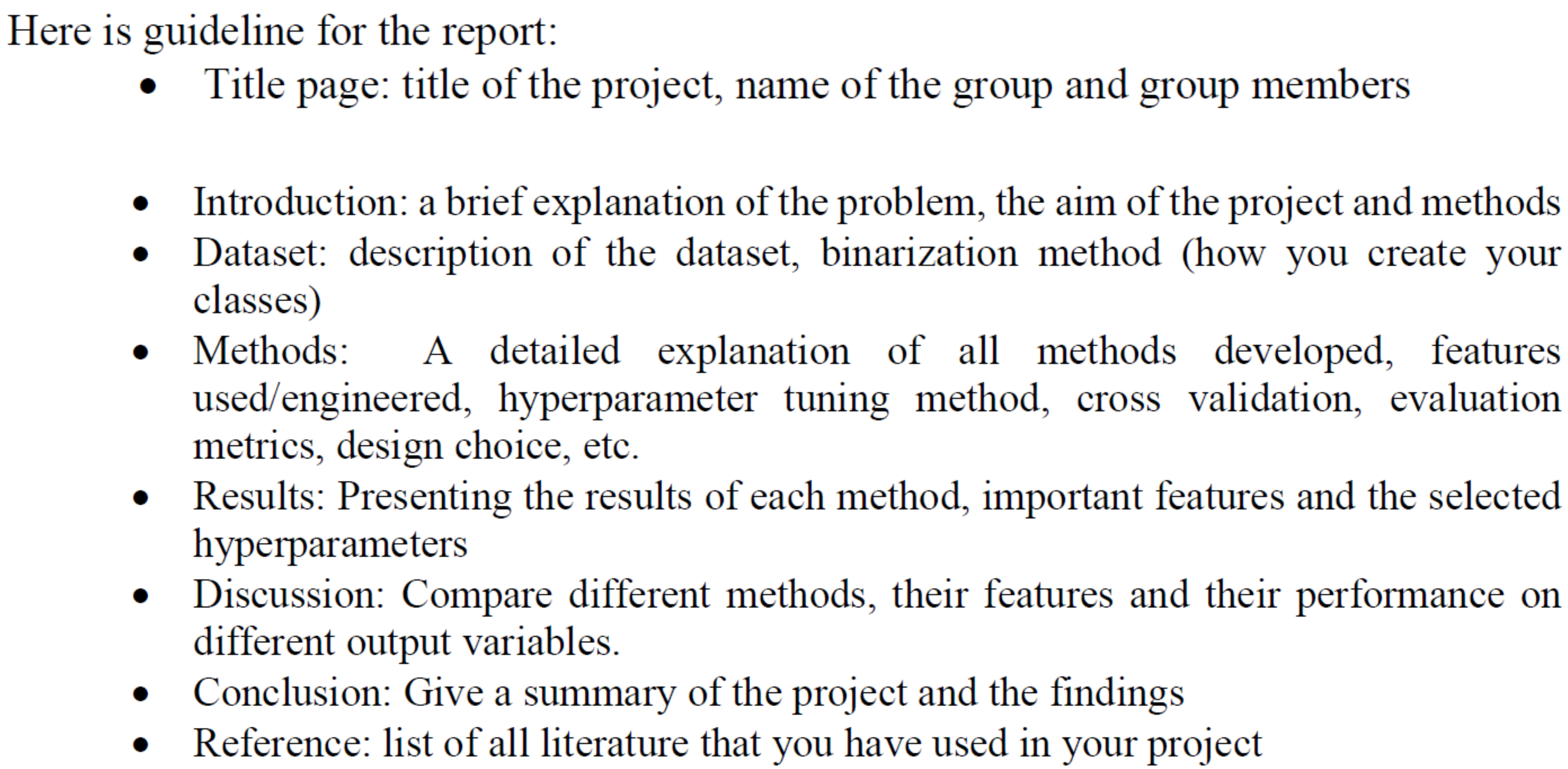


Each group has to submit one report which contains introduction, dataset, methods and evaluation, results, discussion and conclusion. The report is expected to be 12-15 pages (with single column, 1.5 line spacing).

Here is guideline for the report:

- Title page: title of the project, name of the group and group members

Comp9417
Report



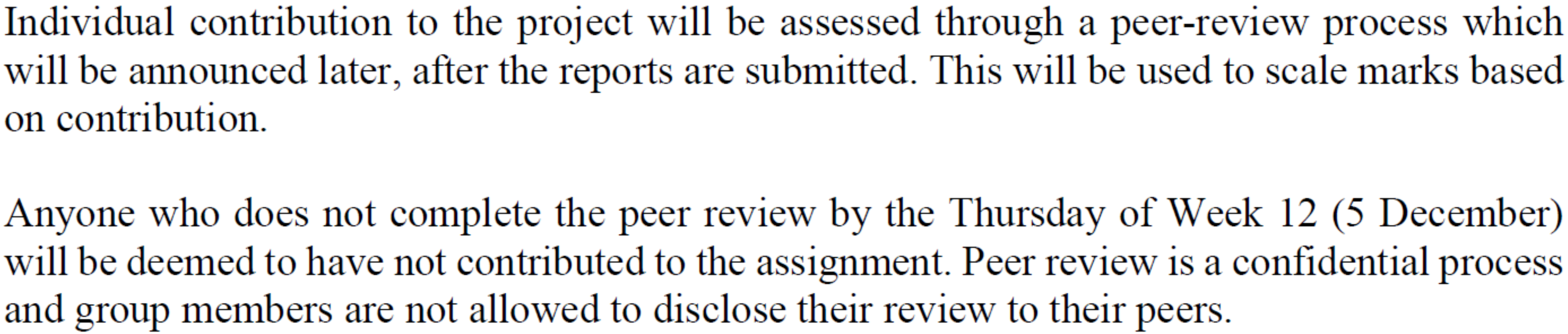
Here is guideline for the report:

- Title page: title of the project, name of the group and group members

- Introduction: a brief explanation of the problem, the aim of the project and methods
- Dataset: description of the dataset, binarization method (how you create your classes)
- Methods: A detailed explanation of all methods developed, features used/engineered, hyperparameter tuning method, cross validation, evaluation metrics, design choice, etc.
- Results: Presenting the results of each method, important features and the selected hyperparameters
- Discussion: Compare different methods, their features and their performance on different output variables.
- Conclusion: Give a summary of the project and the findings
- Reference: list of all literature that you have used in your project

Comp9417

Peer Review



Individual contribution to the project will be assessed through a peer-review process which will be announced later, after the reports are submitted. This will be used to scale marks based on contribution.

Anyone who does not complete the peer review by the Thursday of Week 12 (5 December) will be deemed to have not contributed to the assignment. Peer review is a confidential process and group members are not allowed to disclose their review to their peers.