Internal Report 2012-2013-04 14 June 2013

Universiteit Leiden
Opleiding Informatica & Economie

Self-Tracking with Multiple Sensor Systems

Michiel Vos
Supervisors:
Dr. Arno Knobbe
Ricardo Cachucho MSc.

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Self-tracking is about collecting data from a person's daily life. The goal is to gain better insight into how a person lives and trying to improve the quality of life. Due to innovations and technological progress, sensors systems are becoming more accessible for mainstream users. In this thesis three systems are described, used and combined to collect data of physiology and activity. A dataset and an example how to use it are presented.

Contents

1 Introduction **4**
- 1.1 Preface . . . 4
- 1.2 Motivation . . . 4
- 1.3 Research Questions . . . 4
- 1.4 Outline . . . 4

2 Sensor Systems **6**
- 2.1 BioHarness . . . 6
- 2.2 Beddit . . . 6
- 2.3 OpenBeacon . . . 7

3 Self-Tracking **9**
- 3.1 Data description . . . 9
  - 3.1.1 BioHarness . . . 9
  - 3.1.2 Beddit . . . 9
  - 3.1.3 OpenBeacon . . . 9
- 3.2 Beddit versus BioHarness . . . 10
  - 3.2.1 Comparison . . . 10
  - 3.2.2 Results . . . 11
- 3.3 Data collection . . . 13
- 3.4 Preprocessing . . . 15
  - 3.4.1 From raw to a dataset . . . 15
  - 3.4.2 Extend the dataset . . . 15

4 Feature Construction **17**

5 Data Analysis **18**
- 5.1 CRISP-DM . . . 18
- 5.2 Classification . . . 19
- 5.3 Correlations . . . 20

6 Conclusions and Discussion **23**

Appendices **27**

A Location Model Output **27**

B Data attributes **33**
- B.1 Original . . . 33
- B.2 Raw Dataset . . . 34
- B.3 Dataset used for the correlations of the Data Analysis . . . 35

1 Introduction

1.1 Preface

This bachelor thesis is part of the track Informatica & Economie at the Universiteit Leiden. In the first semester I followed the course Data Mining[1] lectured by Arno Knobbe. I became interested in the subject and wanted to do my bachelor project on one of the projects that were offered by the Data Mining Group. I like to sport and do it at least four times a week. I football at svKMD in the first team and cycle at WTOS, a student club. During cycling I use a Garmin Edge 705 to keep track of my heart rate, speed, average speed and have a spreadsheet with the amount of kilometers per trip. In this project I have combined sport and tracking. Thanks Ricardo Cachucho for assisting me and Arno Knobbe for helping and providing the project.

1.2 Motivation

Actions have consequences and some are clear for us humans and others are not. When something is happening directly after the action and it is consistent, it is easy recognisable as cause-effect. The longer the delay between the cause and effect, the less certain we are about the correlation, but we guess the causality is stronger if the pattern happens repeatedly. For example, drinking coffee before sleep and laying awake in bed later. Still, it is difficult to prove because all other variables that could interrupt sleep should be fixed while experimenting, which is not possible. We could use data mining to help us find causalities about sleep and actions we do during the day. It would give insight how things work and will help improve our life style and quality. In the future, maybe it will be possible to explain health complaints automatically.

1.3 Research Questions

The following questions will be addressed in the thesis. Which devices could be used for self-tracking? Which sensors do they have and what do they measure? What does the data look like they produce? How to collect the data and bundle the data from the different devices with different sampling frequencies? How can the bundled data be used?

1.4 Outline

In the Sensor Systems section, the devices are explained. How do they work, what are the applications, which sensors do they have and what sort of output they produce. In the first part of the Self-Tracking section, is explained what the data means, how the data is structered and if it is possible to extract more data. In the second part, there is a comparison between two sensors. In the third part is explained how all the data is collected from the systems. In the last part of this section the bundling of the data is explained and how to fix the additional problems. In the Feature Construction section is explained how to extract features from the dataset to get a dataset with a different sampling frequency. In the Data Analysis section the CRSIP-DM model is explained and

[1] A definition of data mining is the process of discovering patterns in data, by analysing the data automatically [28]. Another definition is "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [11].

the data is analysed. Multiple models are made to predict the location of the subject given the physiological data of the day. Also, correlations are made, which gives some trivial and some new insights. The last section is about the conclusions, how the self-tracking process could be improved and what the next steps in self-tracking could be.

2 Sensor Systems

2.1 BioHarness

The Zephyr Bioharness 3 [22] is a physiolocigal monitoring device, which is attached to a strap on the chest. There are several practical uses for this device.

- Remote patient monitoring. Medical patients who need health care but want to live at home, in combination with ZephyrLIFE ™[24].
- Athletes who would like to track their progress.
- For coaches to find out who is tired and up for substitution during a match [23].
- In 2010 Chilean mineworkers who were trapped underground were remotedly monitored which resulted in a better rescue order and better health care when the miners were again above ground [25].
- Researchers who could use the data.

The device has an internal storage for more than 500 hours logging and the battery's cycle is up to 35 hours. The device can be connected to a pc with a USB cable, to transfer data and to recharge the battery. The following logs could be produced: general log, summary log, summary and waveform log and the event log. It takes around 1-6 minutes per hour of data to download the log files from the device to a computer [7]. The data is stored in a folder per session (a period from which the device is on till it is turned off) and within the folder different CSV[2] and DaDisp[3] files. The sampling frequency of the CSV files differs from the sampling frequency of the sensors. Most logs are sampled at 1 Hz. The symbol Hz stands for hertz and is defined as the number of cycles per second. The price is $ 472. The accelerometer measures the physical acceleration of the user in the x-, y- and z-axis and its unit is in g. One g is 9.88 metres per second squared. The sampling rate is 100 Hz. The range is from $-16g$ to $16g$ for each axis. The acceleration magnitude is $\sqrt{(\Delta X)^2 + (\Delta Y)^2 + (\Delta Z)^2}$.
The breathing sensor measures the pressure of the chest to the sensor. If the pressure is above a certain threshold, it will count as a breath taken. The breath rate is the amount of breaths taken in a minute. The sampling frequency is 25 Hz and the ranges is from 0 to 120.
The electrocardiography (ECG) [16] sensor measures the electrical activity of the heart. See Figure 2 for a graph of the results that an ECG produces. The sampling frequency is 1000 Hz.

2.2 Beddit

The Beddit Sleep Tracker is placed next to a bed and is connected with a sensor between the sheets and the mattress. The makers of Beddit think sleep is important, because a human is sleeping one third of his live. Better sleep results in a better life quality. Why

[2] CSV stands for Comma-Separated Values and such file can be opend with e.g. Microsoft Excel. Zypher also provides scripts to convert CSV files to Matlab files.

[3] "DADiSP scientific computing and data visualization software that combines the power of programming with the simplicity of a spreadsheet" [8].

Figure 1: The BioHarness.

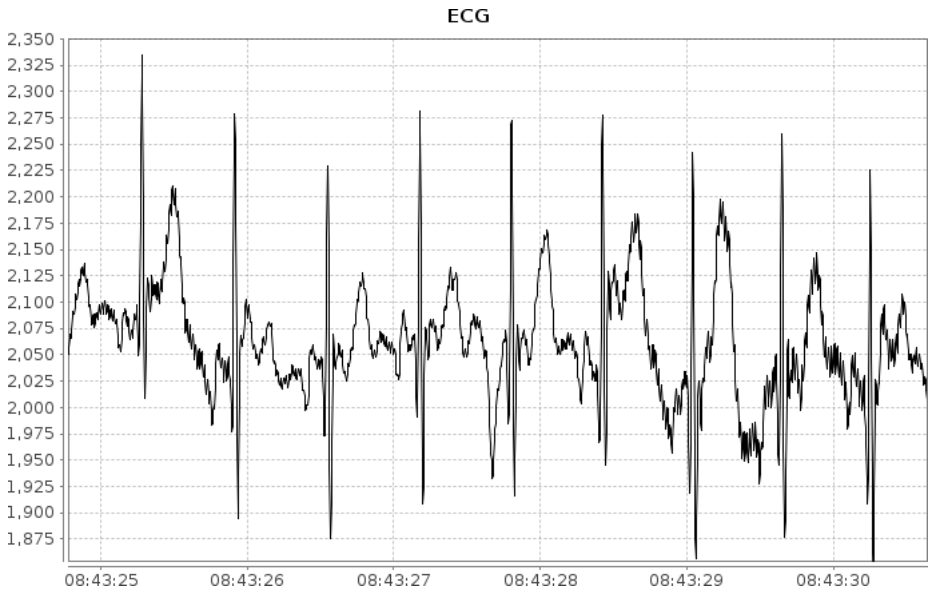

Figure 2: A graph of 5 seconds of ECG data.

not measure it to help improve our sleep quality? The sensor is about 70 cm long and 4 cm wide, so it is possible it will not cover the whole bed. The price is €395.
The ballistocardiogram (BCG) measures (micro)movements of the body. [5] BCG is convenient in use, because you will not notice it is measuring, but the disadvantage compared to ECG it is not only measuring cardiac activity, but also body movements, like tossing and turning, which could also be an advantage. [20] The sampling frequency is 140 Hz. The devices also measure the room temperature (°C), ambient noise level (dB) and brightness (lx), once per 5 minutes [6].
Javascript Object Notation (JSON) is a format to exchange data between different software environments. The data of Beddit is stored in two JSON files per session.

2.3 OpenBeacon

RFID stands for Radio-frequency identification and is a technology to communicate wirelessly between tags. The OpenBeacon Ethernet EasyReader PoE II (called OpenBeacon from now) can receive and send signals with those tags. See Figure 4 for a picture of a tag and the device. There are two kind of tags, active and passive ones. Active tags are powered by a battery and can broadcast their signal. Passive tags do not have batteries and are powered by the energy received wirelessly from the OpenBeacon, but only when

Figure 3: The Beddit sensor and the device [5].

they are in range. The tags of the OpenBeacon are active tags and can also communicate with each other. The OpenBeacon Ethernet EasyReader PoE II can identify RFID tags and edges between the tags. An edge is made between two tags when they are in sight of or near each other. Each tag has it is own ID and each edge has a power level. It takes the OpenBeacon a few seconds to identify the edges. The price of 100 RFID tags is €1575,75 and the EasyReader costs €184,87 euro. RFID tags have a lot of applications, like: theft protection on clothing, as access key pas to enter buildings, track and trace of goods while transporting and social experiments [3] [2]. The OpenBeacon produces JSON output.

Figure 4: A tag for the OpenBeacon [19] and the OpenBeacon Reader.

3 Self-Tracking

3.1 Data description

3.1.1 BioHarness

From the data produced by the sensors, more data can be derived. For the experiments the general log is used, because it has the most interesting variables and a lot of preprocessing is already done. The heart rate is the number of beats per minute and can be derived from the ECG. In the ECG graph (see Figure 2) a pattern is visible, also called a cycle. Each cycle is one heartbeat. The accelerometer measures the acceleration magnitude. The peak acceleration is logged in the general log. The activity level is derived from the accelerometer and is expressed in terms of Vector Magnitude Units (g). The force and direction of gravity is known and the BioHarness is worn at the chest, so it is possible to calculate the posture of the user. The range is 180 degrees. 0° is vertical and 180° is inverted. Breath rate is derived from the breath sensor. The sampling frequency of the general log is 1 Hz.

3.1.2 Beddit

The data from the sensors is being transferred through Wi-Fi or cable to the web servers of Beddit, almost realtime. The servers are analysing the data and extracts the heart rate, respiration (Beddit uses the term respiration for breathing) and the activity from the BCG. When all data is received of the night, the servers are going to analyse it and will compute the sleep stages, sleep efficiency (time sleeping relative to time in bed), average heart rata, average noise level and stress level. The stages are *Away*, *Wake*, *Light sleep*, *Deep sleep*, *REM*[4] and *Missing*. How Beddit computes these stages is not explained, but they probably use the heart rate and the binary actigram. Beddit does not use a constant sampling frequency. For the respiration and the instant heart rate there is a record for every beat, with the BPM computed from two single heart beats and the timestamp in seconds since the start of the session. Presence has a record for every second the presence is changing and a 0 if the user is not in bed and a 1 if the user is in bed. The binary actigram has a record for every second there is a movement above a certain threshold. The minutely actigram has for every minute a value with the amount of movements occurred. The temperature, ambient noise level and brightness have a record every 5 minutes with the date and time in ISO 8601[5] format in local time and their values, respectively celsius, decibel and lux.

3.1.3 OpenBeacon

The OpenBeacon produces JSON, that is stored in a MongoDB database. MongoDB is a NoSQL database, which is the counterpart of a relational database. One argument to use it is because the format of the data is more flexible. The collections of the database can be exported as CSV files. There are two collections. The collection tags, within documents with the format tag ID, timestamp (seconds since the Unix epoch[6]). For every second

[4] REM stands for Rapid Eye Movements, it's a very light sleep and dreaming is possible in this stage.
[5] Example 2013-05-14T19:04 [15]
[6] 00:00:00 GMT January 1, 1970.

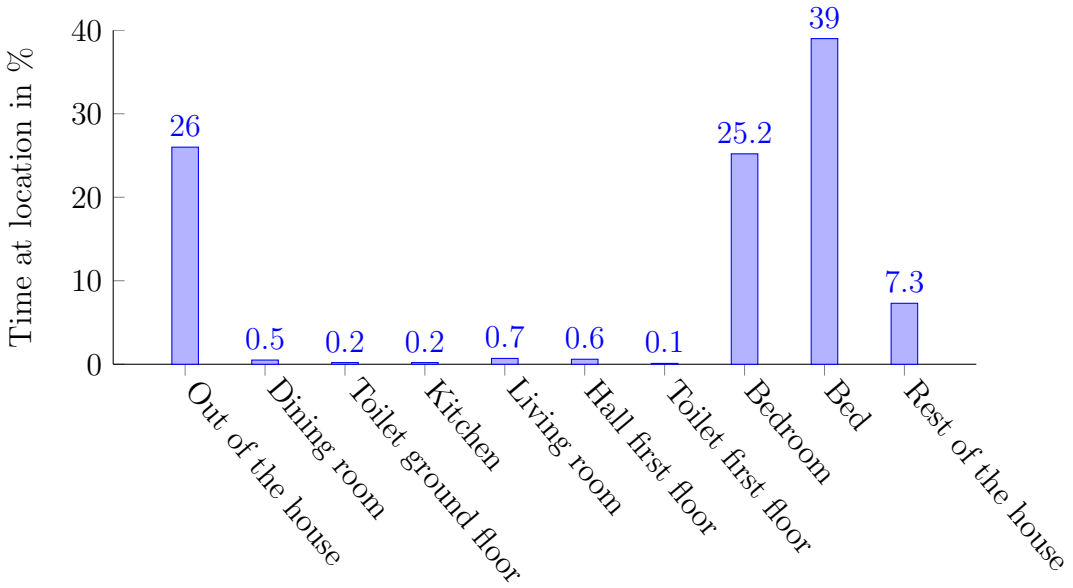

Figure 5: Histogram with the time been at the locations.

the OpenBeacon sees a tag, a document will be added. The second collection is called edges. For every second the OpenBeacon sees an edge between two tags, there will be one document added with the format ID of tag 1, ID of tag 2, power level and timestamp. The higher the power level, the closer the tags are from each other. The power level can not be expressed in terms of distance, because there are other things that could interfere with the power level. For example, walls and other communication on the 2.4 GHz frequency. Figure 5 is a histogram of the locations, with the relative time at the locations. Bed, bedroom and out of the house are the three locations I spend most of my time.

3.2 Beddit versus BioHarness

3.2.1 Comparison

The first experiment is a comparison between the Beddit and the BioHarness. It is a good assignment to get to know the devices and the data. The two devices have overlapping features like the heart rate, breath rate and activity level. It would be a good test to see if they measure the same thing. Beddit does not have a constant sampling frequency. After every heartbeat, the BPM is computed between the heartbeat and the previous heartbeat. The value is notated with the amount of seconds since the session. For every heartbeat, the associated minute and the heart rate is summed up with all the other heartbeat heart rates, and divided by the amount of heartbeats in the specific minute. This results in an average heart rate per minute. The respiration rate is also summed up, but does not need to be divided, because it is the amount of breaths taken. The BioHarness heart rate, breath rate and activity calculations are done in the same way. The difference is that the Bioharness has a constant sampling frequency from 1 Hz. The Beddit binary actigram has already one record per minute, it just needs to be copied. It is not the same as the BioHarness activity, so it would be good to normalise[7] the BioHarness activity and the

[7]Normalisation makes it easier to compare values with different scales.

Beddit binary actigram.

$$\mu = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$D_i = (X_i - \mu)^2$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} D_i}{n}}$$

$$y'_i = \frac{y_i - \bar{y}}{\mu}$$

Figure 6: Z-normalisation [1].

Both devices were used for one night. The sampling frequency is 16.67 mHz (every one minute). Minutes with missing data are removed. Section 3.3 explains more about the setup and how to bundle the data. Sleeping with the BioHarness is possible when laying on the back. Sleeping on the left side is not comfortable. The tool QUICKIE is used to visualise the data and to find the correlations. QUICKIE stands for Quick User Interface for Convolution Kernel-Involving Experiments and is being made and used by the Data Minig Group of Liacs, but is not yet published.

3.2.2 Results

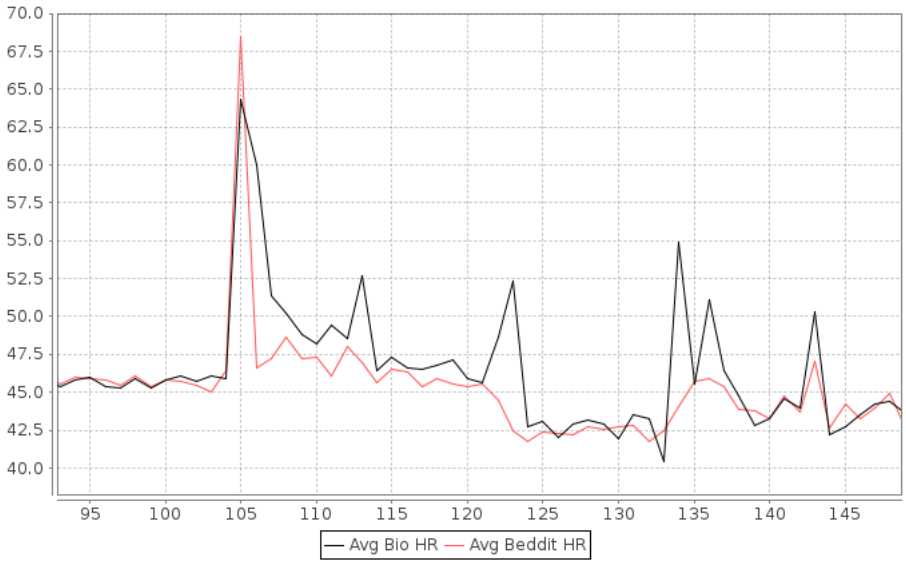

Figure 7: Part of the plot of the heart rate. On the y-axis the heart rate and on the x-axis the amount of minutes after the start of the session.

The lines in Figure 7 are not exactly the same, but they do have the same pattern. The correlation is 0.61. The low correlation could be explained by lag in measurements or one (or both) device(s) does not measure correctly. Moving average is a technique to smooth the lines. If width 7 is used, then it means a value on the new computed line is composed

from the current and the previous 6 minutes of the old line. If moving average is applied to both lines, the correlation of the heart rates will be 0.98.

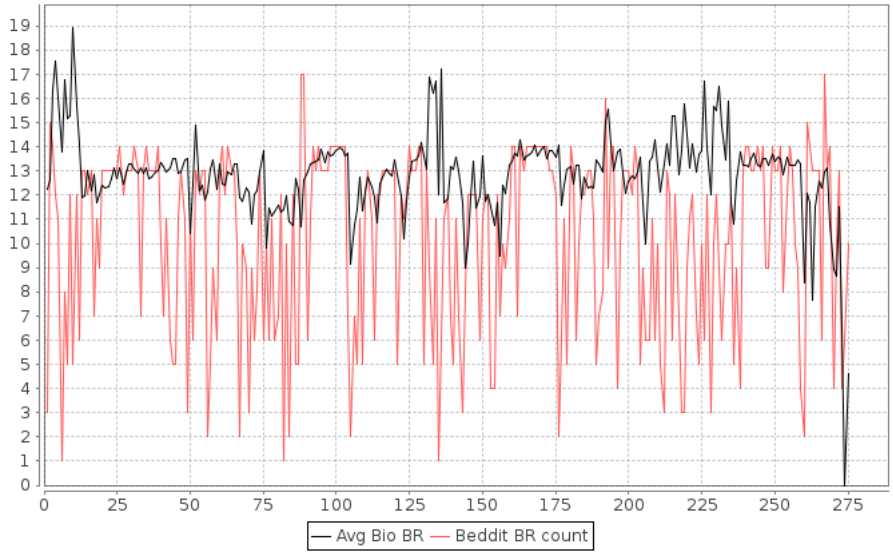

Figure 8: Plot of the breath rate. On the y-axis the heart rate and on the x-axis the amount of minutes after the start of the session.

The lines in Figure 8 are different. The correlation is 0.12. When moving average (width 10) is used the correlation is still 0.71. There is something wrong or the devices are measuring something else. The correlation of the normalised values is 0.26.

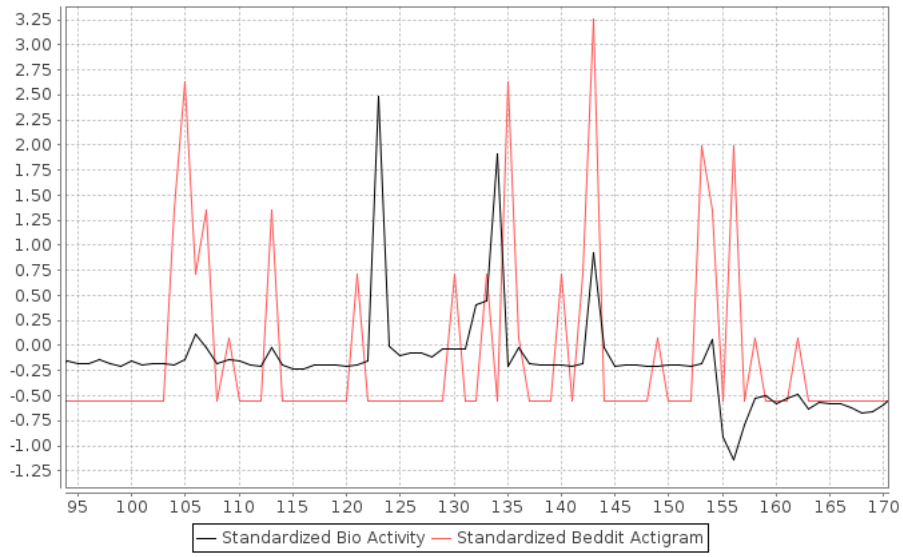

Figure 9: Part of the plot of the activity. On the y-axis the heart rate and on the x-axis the amount of minutes after the start of the session.

The Beddit actigram and the BioHarness activity are something different, so normalisation is used. Figure 9 shows some similarities, but the correlation is still low (0.26). Moving average does not help to improve the correlation.
The heart rates of the Bioharness and the Beddit could be merged into one variable. The BioHarness breath rate and the Beddit respiration rate is totally different. The lines of

the activity are a bit like each other, but the correlation of the activity is low and by far not good enough to merge them into one variable.

3.3 Data collection

Beddit The sensor is placed on my bed and connected with the device. The device is connected to the internet. Beddit starts measuring from 21:30 till 11:00 the next day. After all the measuring, I have downloaded for every night two JSON files with the API of Beddit[6]. After the computing of the data I received an email and I visited the website to see if everything is logged all right.

BioHarness When I wake up, I start my PC and the BioHarness Log Downloader. After about 40 minutes I have took a shower and had breakfast, the Log Downloader is ready. I put the strap on my chest with the device and turn it on. After sport I take a shower and put it off and after shower turn it again on. The device is water resistance up to one meter, but it takes a while for the strap to get dry. Right before I go to bed I connect the device to the PC so it is recharged again the next day. A few minutes of data are missed every day, because of the Log Downloader and the showering.

OpenBeacon In the experiment, the OpenBeacon is used as a localisation tool. The location of the user could explain some behaviour during the day, which can explain behaviour during the night. At first, there were two OpenBeacon devices available, one for the ground floor and one for the second floor. It covered the whole house, but unfortunately one device stopped working. The working device were set up at the ground floor and several tags at the following places: dining room, toilet ground floor, kitchen, living room, hall first floor, toilet first floor. I were also wearing a tag. If there was an edge between my tag and the tag dining room, my location was the dining room for example. The OpenBeacon was always on, but the script to capture the JSON output and to store it in the MongoDB not. After waking, up I started my laptop and the script. I took my tag with me, except in the shower where it was put on a cabinet at the hall. When I left the house the script was stopped and when I came back home I started the script right away. There are a few minutes of missing data every day. For example, when I was late back home from sport and I went to bed, but not before changing clothes, brushing teeth, unpacking my sport bag et cetera, without wearing my tag, because the script was not running. Of course the script is stopped before going to bed.
A few locations were I was at a certain time are known, but more locations could be extracted with some tricks. See Figure 10 for the decision tree.
Strange enough, some tags were showing up which were not part of the tags of the Open-Beacon. These tags were ignored.

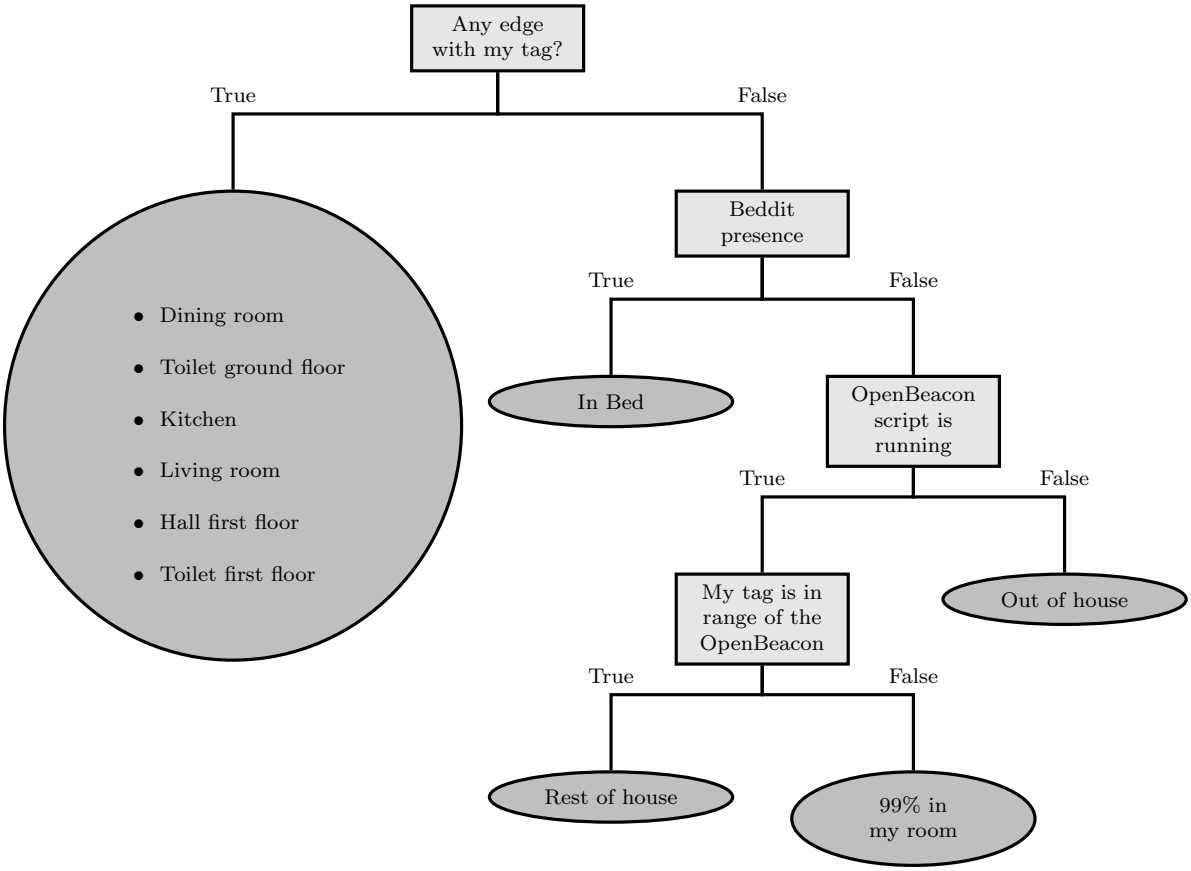

Figure 10: A decision tree how to find more locations.

3.4 Preprocessing

3.4.1 From raw to a dataset

In this phase, all the data is used and combined to have one structured dataset after the phase. The end result is a CSV file. The sampling frequency is 1 Hz, which means for every second a row. The dataset begins at Wed 10 Apr 2013 08:43:01 AM CEST GMT+2 and ends at Thu 25 Apr 2013 08:52:35 AM CEST GMT+2.
Most of the code is loops and more loops, calculating the right timestamp and copy data. The Beddit API only gives records of the sleep stage and timestamp, when the stage is changing. The gaps with missing values are filled in. See Figure 11. There are still a lot of missing values, for example the heart rate at night. It is possible to use methods to fill in the gaps, but it is better to provide a "raw" dataset, with only the facts. When anyone else would like to use the data, they can decide themselves the method or solution.

$$\left(\begin{array}{cc} T & Stage \\ \hline 1 & Away \\ 4 & Wake \\ 6 & Lightsleep \\ 7 & Deepsleep \\ 9 & Lightsleep \\ 11 & Wake \\ 12 & Away \end{array}\right) \rightarrow \left(\begin{array}{cc} T & Stage \\ \hline 1 & Away \\ 2 & Away \\ 3 & way \\ 4 & Wake \\ 5 & Wake \\ 6 & Lightsleep \\ 7 & Deepsleep \\ 8 & Deepsleep \\ 9 & Lightsleep \\ 10 & Lightsleep \\ 11 & Wake \\ 12 & Away \\ 13 & Away \end{array}\right)$$

Figure 11: How the missing values are filled in of the sleep stages.

3.4.2 Extend the dataset

This section gives an example how the missing values can be filled in. The Beddit heart rate gaps are filled in as seen in Figure 12 with linear interpolation. The Beddit heart rate and the BioHarness heart rate are now merged into one variable. There are still missing values in this merged variable and they are also filled in with linear interpolation.

$$\left(\begin{array}{cc} T & HR \\ \hline 1 & 40 \\ 5 & 50 \\ 6 & 52 \\ 9 & 61 \end{array}\right) \rightarrow \left(\begin{array}{cc} T & HR \\ \hline 1 & 40 \\ 2 & 42.5 \\ 3 & 45 \\ 4 & 47.5 \\ 5 & 50 \\ 6 & 52 \\ 7 & 55 \\ 8 & 58 \\ 9 & 61 \\ 10 & 61 \end{array}\right)$$

Figure 12: Linear interpolation to fill in the missing values of the heart rates.

4 Feature Construction

The dataset can be used in many ways and this section describes one way. This section describes a new dataset set based on the extended dataset of Section 3.4.2. This dataset is a set of data which can be used to make a model of the data. Section 5.1 explains more about what a model is. The frequency of the new dataset is on average 11.57 µHz (once every day) and the output is a CSV file with for every day a row. Note a day starts after waking up and it ends the next day before waking up. So not every day is 24 hours long, therefore the sampling frequency is not constant. The heart rate is an example of a feature. It is a feature of the dataset and every second has a heart rate. A new feature is the average heart rate of the day. If the user has a sporty day, the average heart rate will rise. However, it will not rise if the user is going to bed early and has a few hours of extra sleep. The extra sleep will result in a low average heart rate and will compensate the activity during the day. A solution is to split the average heart rate in multiple features. Ranges of heartbeats are selected and the minutes of every range are counted. The new features are *40-*, *40-60*, ... *120-140*, *140+*, *100-*, *100+*. If a minute has an average heart beat of 80, it will lay in two ranges, *80-100* and *100-*. A sporty day will have a lot of minutes in the 100+ range, but still could have the same amount of minutes in the *40-60* range.
Beddit gives in the online interface not only the results of the night, but also gives advice. "To follow good circadian rhythm, try to always wake up at the same time, also in weekends. Irregular sleeping times make it harder for you to fall asleep and to wake up." As response to this advise two features are added; the amount of minutes waking up after 08:00 and the amount of minutes went to bed after 22:00. The values could also be negative, if the wake up or to bed time is before the threshold. In addition to these two features, the difference with the last day is also important to know. How much minutes earlier or later than yesterday did the user wake up or did the user go to bed. During the day, there are ranges of heart rates, but this could also be applied to the night. However, the night is less varied. Beddit can still detect different stages of the sleep (seen in Section 3.1.2). The amount of seconds in a stage are counted. Beddit also provides some extra features about the day, which could not be integrated in the raw dataset of 1 Hz, because of the different sampling frequency. These features are the stress percentage and the resting heart rate and can be integrated in this dataset.
The amount of seconds at the locations are summed up, the average noise, luminosity and temperature are calculated. From the sleep stages the total amount of sleep could be calculated, the amount of wake in bed and also the sleep efficiency (the relative time slept in bed).

5 Data Analysis

5.1 CRISP-DM

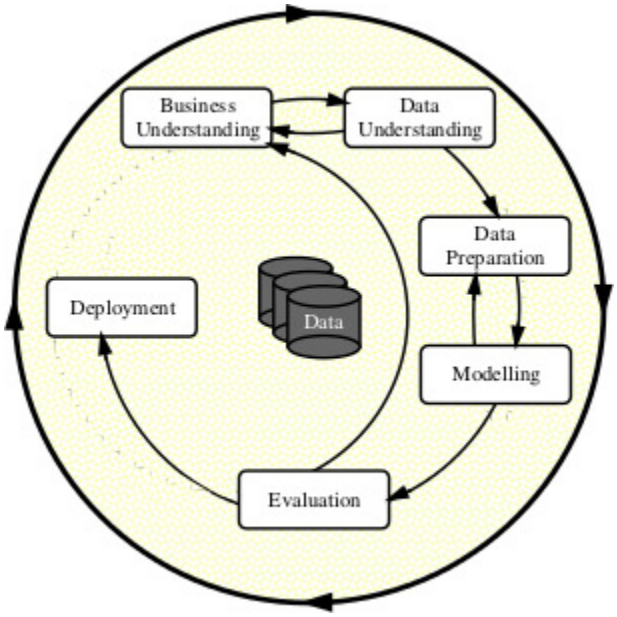

Figure 13: CRISP-DM [27] cycle.

This section has some theory, to get a view of the big picture. Cross Industry Standard Process for Data Mining [27] is invented out of necessity for better planning, documentation and communication for Data Mining projects. It is divided in multiple phases, as seen in Figure 13. The business understanding phase is to describe the urge and the goal of the project, see Section 1.2. The data understanding phase is to learn about the data, what is the source and how it is computed. In Section 2, 3.1 and 3.2 the sensors systems are described, how they are used and what the data is. Normally the data is already collected, but this project needs another phase. The data collection phase is described in Section 3.3. The next phase is the data preparation phase. In this project it exists out of two parts. The first part is about preprocessing to compute the dataset, as seen in Section 3.4. The dataset can be used by the Liacs Data Mining Group for further research. The second part is the feature extraction section, which describes the computing of a new dataset, seen in Section 4. A model is a generalised representation of the data. An algorithm tries to make relations between attributes. There are three reasons to make such a model:

1. Prediction. A bank give loans to clients and they will try to choose the best interest rate, such that the client will accept the interest rate, can pay the interest rate and the loan back over time and the bank will make a profit. Based on data from old clients a model can be made and the probability (prediction) of a client getting problems with paying the intereset rate can be calculated. Based on this information the high of the loan and the interest rate can be determined [14]. Another example is predicting the age of a Twitter user [17].

2. Insight. After analysing the correlation between products bought together in a supermarket, management decided to place diapers and beer close together. More

customers who bought the diapers were tempted to buy the beer just by seeing it, which resulted in more impulsive sales [12].

3. Classification. Email spam (unwanted email) filters are using classification to determine if an email is spam or ham (not spam). Based on the previous received emails with their associated label spam or ham, a model can be made to predict the class of a new received email.

Models need to be trained by a dataset. The bigger the dataset, the better the model, because the chance is higher the set will represent all possible data. The better the model, the better the relations between the attributes are expressed. Section 5.2 makes a decision tree to make it possible to classify the location of a instance based on the data of the BioHarness. Section 5.3 analyzes the correlations between features. The evaluation phase the conclusions are made and will be decided what to do next. The deployment is the implementation to achieve the goal. This could be a new piece of software for bank employees or a new price for a product. In this project there is no deployment.

5.2 Classification

In this section, a model is built of the locations based on the data of the BioHarness and the OpenBeacon. The dataset from Section 3.4.1 is used, with two alterations. The nights are removed, because the BioHarness does not measure in the night. The rows with missing values are also removed. This gives us a dataset with nearly 585000 instances or rows.
WeKa [26] is a software program with a lot of data mining algorithms included and is used to make the model. The used algorithm is J48 and is an implementation of the C4.5 [21] decision tree algorithm. 66% is used as training set and 34% as validation set. At first, leafes with less than 10000 instances are removed. This will result in a pruned tree and will prevent overfitting. As a result of the actions to prevent overfitting, the percentage of correctly classified instances is only 70%, despite the large amount of instances. Only the locations *out*, *bedroom* and *rest of the house* are classified, because those are most visited as seen in Figure 5. The tree is still big with 35 leafes. In this setup the data of the BioHarness could only explain some regular visited locations. It could be improved by using stratifying sampling. This means the same amount of instances are used of every location, but this will make the model less realistic. It could also be improved by using more specific locations, instead of the general *out of the house* location. I could do all kind of activies out of the house, which also applies for rest of the locations. An other solution could be the use of a lower sampling frequency, like $\frac{1}{60}$ Hz (once every minute) instead of 1 Hz (once every second). See Figure 5.2 for the decision tree.
The model is built again, because not all the locations showed up in the desicion tree. The difference is, this time the tree will be pruned less. Only leafes with less than 1000 instances are removed. This results in a larger tree (114 leafes), a higher percentage of correctly classified instances (74%) and more locations showed up (*living.room*). Figure 14 shows the confusion matrix of the second model. The best model possible has only numbers on the diagonal (the green circles), because those are the correctly classified instances. The blue circles show there are many instances classified as *out*, but they were actually *rest.of.the.house* and *bedroom*. The numbers in the red circles were classified as *bedroom*, but were actually *out* and *rest.of.the.house*. These three locations are confused with each

other, but that is logical because those are the major visited places. If you look at the horizontal numbers of *out*, you see the number on the diagonal (from top left to right bottom) is the biggest. This is also true for the horizontal numbers of *bedroom*, but it is not true for *rest.of.the.house* and *living.room*. Most of the instances are predicted wrong for the last two locations.. There are also zero values on the diagonal. This is because the locations *dining.room*, *toilet.ground.floor*, *kitchen*, *hall.first.floor*, *toilet.first.floor* are a small percentage of the dataset and do still not show up in the decision tree.

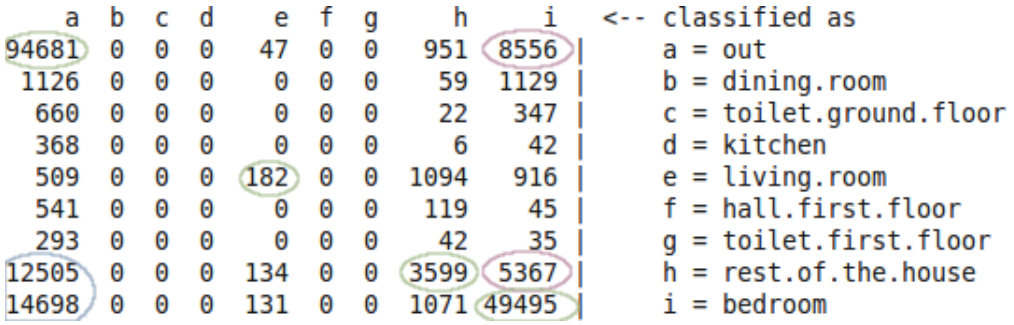

Figure 14: Confusion Matrix.

5.3 Correlations

In Table 1, 28 correlations are listed. These correlations are based on the dataset produced in Section 4. Analysing the results gives some trivial conclusions, but also interesting ones. #3 says when I wake up later, the difference between the previous wake up is greater. The higher the correlations the more consistent the sleep rhythm. #6 says, the later I go to bed, the more I am away from my bed. #11 is also trivial. #1 and #4 are more interesting, because it means when I diverge from my normal sleep rythm, I will sleep less efficiently. Going later to bed and waking up later will result[8] in less REM sleep and it will not be at the expense of deep sleep. As a matter of fact, the correlation is positive, it will result in a bit more deep and light sleep. Going later to bed will result in less time awake in bed, but sleeping longer in the morning will result in more awake time. #18 is also an interesting correlation, but the resting heart rate is relatively constant, so it does not need to mean anything. A high heart rate will result in a less efficient sleep, based on #20 and #27, but based on #21 and #24 it will result in more deep sleep. Most trivial correlation is #28, because stageW is part of the formula to calculate the sleep efficiency.

[8] "Will result" is not precisely true. Correlation does not mean causality, but in this paragraph the causalities are assumed.

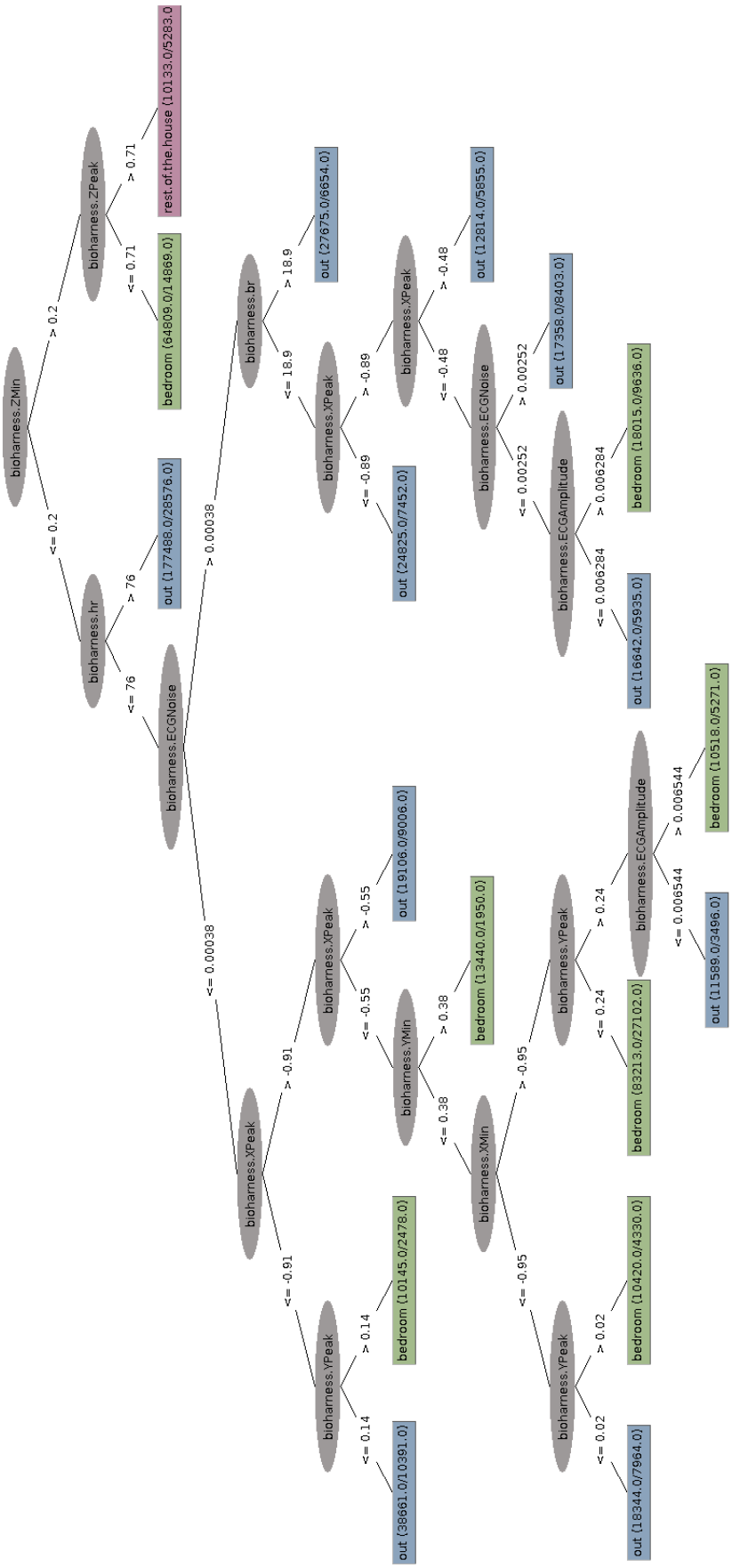

Figure 15: A decision tree of the first model.

#	Feature A	Feature B	Correlation
1	outbed.difference[9]	sleep.efficiency[10]	-0.51
2	outbed[11]	inbed.difference	0.33
3	outbed	outbed.difference	0.54
4	inbed.difference	sleep.efficiency	-0.14
5	inbed[12]	outbed	0.21
6	inbed	stageA[13]	0.50
7	inbed	stageW[14]	-0.53
8	inbed	stageD[15]	0.10
9	inbed	stageL[16]	0.09
10	inbed	stageR[17]	-0.40
11	outbed	stageA	-0.74
12	outbed	stageW	0.13
13	outbed	stageL	0.07
14	outbed	stageD	0.03
15	outbed	stageR	-0.67
16	hr100+	resting.heartrate	0.28
17	stress.percent	resting.heartrate	-0.35
18	out[18]	resting.heartrate	0.65
19	stageD	resting.heartrate	-0.15
20	hr140+	sleep.efficiency	-0.59
21	hr100+	stageD	0.25
22	hr100+	stageL	-0.13
23	hr100+	stageR	-0.41
24	avgHR	stageD	0.32
25	avgHR	stageL	-0.17
26	avgHR	stageR	-0.46
27	avgHR	sleep.efficiency	-0.39
28	stageW	sleep.efficiency	-0.99

Table 1: A few correlations between different features.

[9]The amount of minutes woke up later than the previous day
[10]Relative time slept in bed
[11]Amount of minutes woke up after 08:00 am
[12]Amount of minutes going to bed after 22:00 pm
[13]Away
[14]Wake
[15]Deep sleep
[16]Light sleep
[17]REM sleep
[18]Out of the house

6 Conclusions and Discussion

Working with the three devices is doable for two weeks, but it takes effort to stay focussed and to not forget anything. With more than a million records from multiple data devices, preprocessing takes a lot of time. The code is not advanced, but it is complex because of all the stages the data needs to go through. The way I used the dataset resulted in a new dataset of 15 days, and this is not enough to draw trustworthy conclusions, but it is a start. There are other ways to use the data, as seen in the Classification section, which give more instances. The OpenBeacon is nice to have as an extra, but in my situation a location could not explain proper the things I do, because I use the locations for multiple things and most of the time I am in my bedroom. Still, the decision tree could predict 74% of the instances correctly.
There are more interesting subjects than me for the experiment. I live structurally, do not have health complains, do not smoke, do not do drugs, drink no or little alcohol, drink no coffee and sport a lot. Not a typical student. When a subject is chosen with an irregular live schedule, the correlations between the different features could be different, because relations could turn up more visible. Before the experiment, I was thinking about doing something in one week and doing something else the other week. It would be interesting to see how I would react. For example drinking coffee before going to bed. However, data mining has a paradigm to not use hypotheses. If there is a correlation, it will show up eventually.
Despite the few days of data, I could agree with the correlations measered. Going later to bed and waking up later will result in more deep sleep, more light sleep and less REM sleep. Based on this, I would advise myself to sleep less. With an average of 9 hours and 24 minutes in bed per night, I think that is fair. The interface of Beddit is easy to use and also informative. An advice I read on the website of Beddit was to sleep more structural, even more structurally then I already did. Now I am trying to wake up and go to bed every day at the same time. Beddit is still running at the time of writing this and I visit the website of Beddit every day. The statistics motivate me to keep me to my schedule and to improve my life style. I changed the environment to make it darker and sometimes use earplugs to ignore noise.
In the future more devices could be added or replaced to the experiment. The Nike FuelBand [18] is an activity tracker comparable with the BioHarness. It has an accelerometer, tracks each step taken and also the amount of burned calories. In the next version, it will also track the heart rate [9]. It is more convenient to use than the BioHarness, because it is worn on the wrist. Measuring 24/7 will be posible. The OpenBeacon is only able to find my location in a part of the house. With the help of a smartphone and software it is possible to track someone's location more precisely. GPS could be used to get a position outside and Wi-Fi could be used to get a position in an indoor environment [13]. Fitbit Aria [10] is a scale and it tracks the weight, body fat percentage and the BMI (Body Mass Index). It will also help the user to log the meals and activity. If these devices are added to the experiment, with a better subject and for a longer time measurement, more insight will be gained.
In 2-4 years, more and more wearable sensors systems are going to be used by the mainstream. Nike already started the trend with the Nike FuelBand, but was certainly not the first device in its kind. They have a better marketing system and can reach normal people. Due to economies of scale, they can produce it for a better price. The product has

a small network externality[19]. The users of the Nike FualBand like to compare each other and also themselves with top athletes. It also motivates each other to get active. Most of the devices will publish an API[20]. I predict a social community which will bundle all the APIs, will present all the data from the different devices in a user-friendly way and make it shareable. By that time it will interesting to ask the owner of the social community to research the data and do data mining experiments with it. In this project one user is measured for 15 days with 3 devices. Imagine a dataset with 1000+ users with more devices, measered for over a year.
The car insurance company Axa is building a smart phone application which could check behaviour of the driver [4]. The application will give the drivers advice based on GPS and weather data. The company is researching if it is possible to give the drivers a discount on the premium by good behaviour. I think the next generation of cars could also be connected to a smart phone and more data is available for research. This is just for a driver and his car, but it could also be applied to health insurance. There will be privacy concerns, but eventually customers will take the discount in exchange for their privacy.
This project is a first approach to use self-tracking with multiple sensor devices. When the demand for these wearable self-tracking sensors is increased, the process of self-tracking will get more efficient and more automated. The problem of having not enough data is solved and analysing the data will be more interesting.

[19] A phone has also a network externality. The very first user of a phone had no use of it. He could not call anyone else.

[20] An Application Programming Interface makes it possible to use the data by external developers.

References

[1] A Alan and B Finlay. *Statistical Methods for the Social Sciences*. 4th. Pearson International Edition, 2009. ISBN: 0-13-713150-X.

[2] M Atzmüeller et al. "Enhancing Social Interactions at Conferences". In: *Informatik-Spektrum* ().

[3] A Barrat et al. "High Resolution Dynamical Mapping of Social Interactions With Active RFID". In: *ArXiv e-prints* (Nov. 2008).

[4] K Baumers. *Autoverzekeraar controleert rijgedrag via smartphone*. [Online; accessed 03-June-2013]. URL: `http://www.nieuwsblad.be/article/detail.aspx?articleid=DMF20130531_00606449`.

[5] Beddit. *Science behind Beddit*. [Online; accessed 19-May-2013]. URL: `http://beddit.com/science`.

[6] *Beddit API v2*. [Online; accessed 19-May-2013]. Beddit. URL: `https://docs.google.com/document/d/1D-JULr_zu_B80wFnNgyNjPNPTZXquqm6cqwyRcDUHas/pub`.

[7] *BioHarness 3 Data Sheet*. [Online; accessed 19-May-2013]. Zephyr Technology. 2012. URL: `http://www.zephyranywhere.com/media/pdf/BioHarness_3-DataSheet-2012-JUL-13.pdf`.

[8] *DADiSP - The Ultimate Engineering Spreadsheet*. [Online; accessed 21-May-2013]. URL: `http://dadisp.com`.

[9] C Davies. *Nike FuelBand 2 reportedly ads heart rate monitor and BT 4.0*. [Online; accessed 03-June-2013]. URL: `http://www.slashgear.com/nike-fuelband-2-reportedly-adds-heart-rate-monitor-and-bt-4-0-08280987/`.

[10] Fitbit. *Fitbit Aria Wi-Fi Smart Scale*. [Online; accessed 03-June-2013]. URL: `http://www.fitbit.com/aria`.

[11] W Frawley, G Piatetsky-Shapiro, and C Matheus. "Knowledge Discovery in Databases: An Overview". In: *Al Magazine* (1992).

[12] G Fuchs. *Data mining - If Only It Really Were about Beer and Diapers*. [Online; accessed 30-May-2013]. 2004. URL: `http://www.information-management.com/news/1006133-1.html`.

[13] A Howard, S Siddiqi, and G Sukhatme. "An Experimental Study of Localization Using Wireless Ethernet". In: *4th International Conference on Field and Service Robotics* (2003).

[14] C Huang, M Chen, and C Wang. *Credit scoring with a data mining approach based on support vector machines*.

[15] ISO. *ISO 8601: Data elements and interchange formats*. [Online; accessed 21-May-2013]. 2004. URL: `http://dotat.at/tmp/ISO_8601-2004_E.pdf`.

[16] S Meek and F Morris. "ABC of clinical electrocardiography". In: *BMJ* 324 (2002).

[17] D Nguyen et al. "'How Old Do You Think I Am?': A Study of Language and Age in Twitter'". In: *Seventh International AAAI Conference on Weblogs and Social Media* (2013). URL: `http://www.dongnguyen.nl/publications/nguyen-icwsm2013.pdf`.

[18] Nike. *Nike+ FuelBand. Tracks your all-day activity and helps you do more.*. [Online; accessed 02-June-2013]. URL: http://www.nike.com/us/en_us/c/nikeplus-fuelband.

[19] *OpenBeacon Active RFID Project*. [Online; accessed 27-May-2013]. URL: http://www.openbeacon.org.

[20] J Paalasmaa and M Ranta. *Detecting Heartbeats in the Ballistocardiogram with Clustering*. [Online; accessed 19-May-2013]. 2008. URL: http://www.cs.helsinki.fi/u/jpaalasm/paalasmaa_2008_detecting.pdf.

[21] John Ross Quinlan. *C4. 5: programs for machine learning*. Vol. 1. Morgan kaufmann, 1993.

[22] Zephyr Technology. *BioHarness 3*. [Online; accessed 10-May-2013]. 2013. URL: http://www.zephyr-technology.com/products/bioharness-3/.

[23] Zephyr Technology. *Coaches & Athletes*. [Online; accessed 18-May-2013]. URL: http://www.zephyranywhere.com/training-systems/coaches-athletes/.

[24] Zephyr Technology. *Patient Centrix Monitoring Throughout the Continuum of Clinical Care*. [Online; accessed 18-May-2013]. URL: http://www.zephyranywhere.com/media/WhitePapers/WhitePaper-ZWP-008-Zephy-Backgrounder-with-description-of-Care-Beyond-Walls-Project_For%20Zephyr_Final.pdf.

[25] Zephyr Technology. *Zephyr Provides Physiological Monitoring of Chilean Miners During San Jose Min Rescue Operation*. [Online; accessed 18-May-2013]. URL: http://www.zephyranywhere.com/media/CaseStudies/ZCS-007-CaseStudy-HC_ChileanMinerRescueOperation.pdf.

[26] Weka. *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [Online; accessed 11-June-2013]. URL: http://www.cs.waikato.ac.nz/ml/weka/.

[27] R Wirth and J Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (2000), pp. 29–39.

[28] I Witten and E Frank. *Data mining: Practical Machine Learning Tools and Techniques*. 2nd. Diane Cerra or Morgan Kaufmann publishers or Elsevier, 2005. ISBN: 0-12-088407-0.

Appendices

A Location Model Output

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 1000
Relation:     self-tracking
Instances:    584619
Attributes:   15
              hr
              br
              activity
              acceleration
              posture
              BRAmplitude
              ECGAmplitude
              ECGNoise
              XMin
              XPeak
              YMin
              YPeak
              ZMin
              ZPeak
              location
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

ZMin <= 0.2
|  hr <= 76
|  |  ECGNoise <= 0.00038
|  |  |  XPeak <= -0.91
|  |  |  |  YMin <= 0.15
|  |  |  |  |  ZMin <= -0.12
|  |  |  |  |  |  hr <= 55
|  |  |  |  |  |  |  ZMin <= -0.16
|  |  |  |  |  |  |  |  ECGNoise <= 0.00014: out (1033.0/465.0)
|  |  |  |  |  |  |  |  ECGNoise > 0.00014: bedroom (3476.0/1421.0)
|  |  |  |  |  |  |  ZMin > -0.16: out (2877.0/950.0)
|  |  |  |  |  |  hr > 55: out (12380.0/3873.0)
|  |  |  |  |  ZMin > -0.12: out (20681.0/3495.0)
|  |  |  |  YMin > 0.15: bedroom (8342.0/1098.0)
|  |  |  XPeak > -0.91
|  |  |  |  XPeak <= -0.55
|  |  |  |  |  YMin <= 0.38
```

```
| | | | | | ECGAmplitude <= 0.007764
| | | | | | | BRAmplitude <= 36
| | | | | | | | ECGAmplitude <= 0.004804: out (2055.0/493.0)
| | | | | | | | ECGAmplitude > 0.004804
| | | | | | | | | hr <= 59
| | | | | | | | | | XPeak <= -0.84
| | | | | | | | | | | YMin <= -0.06
| | | | | | | | | | | | ECGNoise <= 0.00014: out (1147.0/397.0)
| | | | | | | | | | | | ECGNoise > 0.00014
| | | | | | | | | | | | | hr <= 52: bedroom (1048.0/259.0)
| | | | | | | | | | | | | hr > 52: out (1182.0/589.0)
| | | | | | | | | | | YMin > -0.06: bedroom (2088.0/184.0)
| | | | | | | | | | XPeak > -0.84: bedroom (18068.0/4166.0)
| | | | | | | | | hr > 59
| | | | | | | | | | BRAmplitude <= 18: out (1123.0/156.0)
| | | | | | | | | | BRAmplitude > 18
| | | | | | | | | | | YMin <= 0.03
| | | | | | | | | | | | posture <= 22: out (1218.0/348.0)
| | | | | | | | | | | | posture > 22
| | | | | | | | | | | | | ECGAmplitude <= 0.006924: bedroom (1121.0/385.0)
| | | | | | | | | | | | | ECGAmplitude > 0.006924: out (1393.0/718.0)
| | | | | | | | | | | YMin > 0.03: bedroom (1801.0/648.0)
| | | | | | | BRAmplitude > 36
| | | | | | | | YMin <= 0.21
| | | | | | | | | XMin <= -0.93
| | | | | | | | | | ZMin <= 0.14
| | | | | | | | | | | ZMin <= -0.15
| | | | | | | | | | | | br <= 20.1
| | | | | | | | | | | | | acceleration <= 0.14
| | | | | | | | | | | | | | XMin <= -0.97: out (1153.0/494.0)
| | | | | | | | | | | | | | XMin > -0.97
| | | | | | | | | | | | | | | hr <= 57: out (1539.0/598.0)
| | | | | | | | | | | | | | | hr > 57: bedroom (1874.0/1016.0)
| | | | | | | | | | | | | acceleration > 0.14
| | | | | | | | | | | | | | YMin <= -0.16: bedroom (1051.0/524.0)
| | | | | | | | | | | | | | YMin > -0.16: out (1275.0/712.0)
| | | | | | | | | | | | br > 20.1: out (1009.0/319.0)
| | | | | | | | | | | ZMin > -0.15: out (3977.0/1057.0)
| | | | | | | | | | ZMin > 0.14: bedroom (1355.0/556.0)
| | | | | | | | | XMin > -0.93
| | | | | | | | | | YMin <= -0.15: bedroom (8815.0/2214.0)
| | | | | | | | | | YMin > -0.15
| | | | | | | | | | | posture <= 44
| | | | | | | | | | | | br <= 15.3
| | | | | | | | | | | | | hr <= 56
| | | | | | | | | | | | | | XMin <= -0.84: out (1029.0/406.0)
| | | | | | | | | | | | | | XMin > -0.84: bedroom (1046.0/552.0)
| | | | | | | | | | | | | hr > 56
| | | | | | | | | | | | | | BRAmplitude <= 62: rest.of.the.house (1673.0/1200.0)
| | | | | | | | | | | | | | BRAmplitude > 62: bedroom (1113.0/706.0)
```

```
| | | | | | | | | | | | | br > 15.3: bedroom (7205.0/3155.0)
| | | | | | | | | | | | posture > 44: bedroom (1933.0/468.0)
| | | | | | | | | YMin > 0.21
| | | | | | | | | | ZMin <= -0.33: out (13322.0/4156.0)
| | | | | | | | | | ZMin > -0.33: bedroom (1230.0/67.0)
| | | | | | | ECGAmplitude > 0.007764
| | | | | | | | posture <= -24: living.room (1386.0/752.0)
| | | | | | | | posture > -24
| | | | | | | | | posture <= 21
| | | | | | | | | | ZPeak <= -0.08
| | | | | | | | | | | hr <= 55: bedroom (1156.0/442.0)
| | | | | | | | | | | hr > 55: out (2365.0/674.0)
| | | | | | | | | | ZPeak > -0.08: bedroom (1044.0/629.0)
| | | | | | | | | posture > 21
| | | | | | | | | | BRAmplitude <= 23
| | | | | | | | | | | ZMin <= -0.65
| | | | | | | | | | | | hr <= 54
| | | | | | | | | | | | | ECGAmplitude <= 0.008904: bedroom (1437.0/548.0)
| | | | | | | | | | | | | ECGAmplitude > 0.008904: out (1741.0/789.0)
| | | | | | | | | | | | hr > 54: bedroom (1157.0/652.0)
| | | | | | | | | | | ZMin > -0.65: bedroom (6882.0/1709.0)
| | | | | | | | | | BRAmplitude > 23: bedroom (34053.0/9614.0)
| | | | | YMin > 0.38
| | | | | | ZMin <= -0.49: out (1289.0/576.0)
| | | | | | ZMin > -0.49: bedroom (12151.0/942.0)
| | | | XPeak > -0.55
| | | | | YPeak <= 0.2
| | | | | | BRAmplitude <= 29
| | | | | | | ECGNoise <= 0.0002: bedroom (1085.0/533.0)
| | | | | | | ECGNoise > 0.0002: out (3235.0/995.0)
| | | | | | BRAmplitude > 29
| | | | | | | hr <= 58: bedroom (2617.0/633.0)
| | | | | | | hr > 58: out (1401.0/711.0)
| | | | | YPeak > 0.2
| | | | | | BRAmplitude <= 33
| | | | | | | hr <= 52: out (1111.0/361.0)
| | | | | | | hr > 52: rest.of.the.house (1570.0/867.0)
| | | | | | BRAmplitude > 33
| | | | | | | br <= 14.8
| | | | | | | | XPeak <= -0.39: rest.of.the.house (1879.0/781.0)
| | | | | | | | XPeak > -0.39: out (1061.0/365.0)
| | | | | | | br > 14.8: out (5078.0/1098.0)
| | ECGNoise > 0.00038
| | | posture <= -18
| | | | posture <= -27: bedroom (1436.0/485.0)
| | | | posture > -27
| | | | | YPeak <= -0.16: out (2406.0/679.0)
| | | | | YPeak > -0.16: bedroom (1590.0/335.0)
| | | posture > -18
| | | | ECGAmplitude <= 0.005224: out (28147.0/6249.0)
```

```
| | | | ECGAmplitude > 0.005224
| | | | | br <= 18.2
| | | | | | XPeak <= -0.89
| | | | | | | YMin <= -0.02: out (12554.0/3919.0)
| | | | | | | YMin > -0.02
| | | | | | | | YMin <= 0.04: out (1071.0/517.0)
| | | | | | | | YMin > 0.04: bedroom (1014.0/558.0)
| | | | | | XPeak > -0.89
| | | | | | | posture <= 57
| | | | | | | | ECGNoise <= 0.00174
| | | | | | | | | ECGAmplitude <= 0.006304
| | | | | | | | | | YPeak <= 0.26
| | | | | | | | | | | XMin <= -0.89: out (1893.0/951.0)
| | | | | | | | | | | XMin > -0.89: bedroom (1423.0/736.0)
| | | | | | | | | | YPeak > 0.26: out (4057.0/1240.0)
| | | | | | | | | ECGAmplitude > 0.006304
| | | | | | | | | | YPeak <= 0.28
| | | | | | | | | | | ZPeak <= -0.32
| | | | | | | | | | | | ZMin <= -0.77: out (1098.0/485.0)
| | | | | | | | | | | | ZMin > -0.77: bedroom (7000.0/3016.0)
| | | | | | | | | | | ZPeak > -0.32
| | | | | | | | | | | | ZPeak <= -0.14: out (2533.0/1020.0)
| | | | | | | | | | | | ZPeak > -0.14
| | | | | | | | | | | | | BRAmplitude <= 79: out (1477.0/827.0)
| | | | | | | | | | | | | BRAmplitude > 79: bedroom (1033.0/583.0)
| | | | | | | | | | YPeak > 0.28
| | | | | | | | | | | XPeak <= -0.66: bedroom (1518.0/611.0)
| | | | | | | | | | | XPeak > -0.66: out (1079.0/586.0)
| | | | | | | | ECGNoise > 0.00174
| | | | | | | | | YMin <= -0.15
| | | | | | | | | | activity <= 0.18: out (5126.0/2417.0)
| | | | | | | | | | activity > 0.18
| | | | | | | | | | | BRAmplitude <= 77: out (1729.0/983.0)
| | | | | | | | | | | BRAmplitude > 77: rest.of.the.house (2273.0/1420.0)
| | | | | | | | | YMin > -0.15
| | | | | | | | | | BRAmplitude <= 55: bedroom (2575.0/1480.0)
| | | | | | | | | | BRAmplitude > 55: out (4676.0/2039.0)
| | | | | | | posture > 57
| | | | | | | | ZMin <= -0.75
| | | | | | | | | posture <= 76: out (3273.0/979.0)
| | | | | | | | | posture > 76: rest.of.the.house (1010.0/553.0)
| | | | | | | | ZMin > -0.75: rest.of.the.house (1109.0/485.0)
| | | | | br > 18.2
| | | | | | ECGAmplitude <= 0.009024
| | | | | | | posture <= 18: out (9145.0/2127.0)
| | | | | | | posture > 18
| | | | | | | | XPeak <= -0.57
| | | | | | | | | br <= 21.7
| | | | | | | | | | hr <= 56: bedroom (1376.0/614.0)
| | | | | | | | | | hr > 56: out (2913.0/1339.0)
```

```
| | | | | | | | | br > 21.7: out (1105.0/454.0)
| | | | | | | | | XPeak > -0.57: out (3814.0/1090.0)
| | | | | | | ECGAmplitude > 0.009024: out (5738.0/872.0)
| | hr > 76
| | | hr <= 142
| | | | posture <= 8
| | | | | ZPeak <= 0.4
| | | | | | ECGAmplitude <= 0.006264: out (4731.0/1105.0)
| | | | | | ECGAmplitude > 0.006264
| | | | | | | ZPeak <= 0.08: out (1028.0/356.0)
| | | | | | | ZPeak > 0.08: rest.of.the.house (1697.0/1013.0)
| | | | | ZPeak > 0.4: rest.of.the.house (1089.0/489.0)
| | | | posture > 8: out (125526.0/23531.0)
| | | hr > 142: out (43088.0/1065.0)
ZMin > 0.2
| ZMin <= 0.8
| | YMin <= 0.28
| | | hr <= 62
| | | | posture <= -28
| | | | | ECGAmplitude <= 0.007864
| | | | | | ZMin <= 0.62
| | | | | | | ECGAmplitude <= 0.005244
| | | | | | | | posture <= -49: bedroom (2192.0/36.0)
| | | | | | | | posture > -49
| | | | | | | | | ZPeak <= 0.34: bedroom (1681.0/200.0)
| | | | | | | | | ZPeak > 0.34
| | | | | | | | | | ECGNoise <= 0.00016
| | | | | | | | | | | BRAmplitude <= 38: rest.of.the.house (1023.0/343.0)
| | | | | | | | | | | BRAmplitude > 38: living.room (1270.0/682.0)
| | | | | | | | | | ECGNoise > 0.00016: bedroom (1090.0/415.0)
| | | | | | | | ECGAmplitude > 0.005244: bedroom (21125.0/2963.0)
| | | | | | | ZMin > 0.62
| | | | | | | | YMin <= 0.08
| | | | | | | | | ECGNoise <= 0.00012: rest.of.the.house (1078.0/442.0)
| | | | | | | | | ECGNoise > 0.00012: bedroom (1092.0/439.0)
| | | | | | | | YMin > 0.08: bedroom (1002.0/55.0)
| | | | | | ECGAmplitude > 0.007864: rest.of.the.house (1597.0/935.0)
| | | | | posture > -28
| | | | | | posture <= -18
| | | | | | | ECGAmplitude <= 0.005024
| | | | | | | | ECGNoise <= 0.00026: bedroom (2171.0/377.0)
| | | | | | | | ECGNoise > 0.00026: out (1075.0/359.0)
| | | | | | | ECGAmplitude > 0.005024: bedroom (21854.0/2103.0)
| | | | | | posture > -18
| | | | | | | YMin <= -0.16: out (1034.0/445.0)
| | | | | | | YMin > -0.16: bedroom (1348.0/605.0)
| | | | hr > 62
| | | | | ZMin <= 0.61
| | | | | | activity <= 0.06
| | | | | | | ZMin <= 0.26
```

```
| | | | | | | BRAmplitude <= 78: out (1842.0/845.0)
| | | | | | | BRAmplitude > 78: bedroom (1197.0/412.0)
| | | | | | ZMin > 0.26: bedroom (2312.0/593.0)
| | | | | activity > 0.06: bedroom (1064.0/621.0)
| | | | ZMin > 0.61: out (1189.0/194.0)
| | YMin > 0.28
| | | ECGNoise <= 0.0001: rest.of.the.house (1508.0/684.0)
| | | ECGNoise > 0.0001: bedroom (1063.0/421.0)
| ZMin > 0.8: rest.of.the.house (4132.0/1086.0)

Number of Leaves  :  114

Size of the tree :  227


Time taken to build model: 23249.86 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      147957               74.4363 %
Incorrectly Classified Instances     50813               25.5637 %
Kappa statistic                      0.5404
Mean absolute error                  0.0853
Root mean squared error              0.2067
Relative absolute error              3.5111 %
Root relative squared error          9.7501 %
Total Number of Instances           198770

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.908     0.325       0.755    0.908      0.825      0.871    out
                   0         0           0        0          0          0.821    dining.room
                   0         0           0        0          0          0.852    toilet.ground.floor
                   0         0           0        0          0          0.831    kitchen
                   0.067     0.002       0.368    0.067      0.114      0.91     living.room
                   0         0           0        0          0          0.841    hall.first.floor
                   0         0           0        0          0          0.819    toilet.first.floor
                   0.167     0.019       0.517    0.167      0.252      0.711    rest.of.the.house
                   0.757     0.123       0.751    0.757      0.754      0.889    bedroom
Weighted Avg.      0.744     0.213       0.704    0.744      0.709      0.859

=== Confusion Matrix ===

     a   b   c   d    e   f   g    h     i   <-- classified as
 94681   0   0   0   47   0   0  951  8556 |     a = out
  1126   0   0   0    0   0   0   59  1129 |     b = dining.room
   660   0   0   0    0   0   0   22   347 |     c = toilet.ground.floor
```

```
  368  0  0  0    0  0  0     6    42 |     d = kitchen
  509  0  0  0  182  0  0  1094   916 |     e = living.room
  541  0  0  0    0  0  0   119    45 |     f = hall.first.floor
  293  0  0  0    0  0  0    42    35 |     g = toilet.first.floor
12505  0  0  0  134  0  0  3599  5367 |     h = rest.of.the.house
14698  0  0  0  131  0  0  1071 49495 |     i = bedroom
```

B Data attributes

B.1 Original

BioHarness	Beddit	OpenBeacon
Timestamp	Stress percent	Time
Heart rate	Time in bed	Tag A
Breath rate	Time light sleep	Tag B
Posture	Time sleeping	Power level
Activity	Time deep sleep	
Acceleration	Sleep efficiency	
Battery	Resting heartrate	
Breath rate amplitude	Minutely actigram	
ECG amplitude	Noise	
ECG noise	Luminosity	
XMin	Temperature	
XPeak	Sleep stage	
YMin	Respiration	
YPeak	Presence	
ZMin	Instant heart rate	
ZPeak	Binary actigram	

B.2 Raw Dataset

Timestamp
Out
Dining room
Toilet ground floor
Kitchen
Living room
Hall first floor
Toilet first floor
Rest of the house
Bedroom
Bed
Beddit respiration
Beddit respiration minToMin
Beddit respiration maxToMax
Beddit respiration amplitude
Beddit instant heart rate
Beddit sleep stage
Beddit noise
Beddit luminosity
Beddit temperature
Beddit minutely actigram
BioHarness heart rate
BioHarness breath rate
BioHarness activity
BioHarness acceleration
BioHarness posture
BioHarness breath rate amplitude
BioHarness ECG amplitude
Bioharness ECG noise
BioHarness XMin
BioHarness XPeak
BioHarness YMin
BioHarness YPeak
BioHarness ZMin
BioHarness ZPeak

B.3 Dataset used for the correlations of the Data Analysis

Day
Heart rate average
Out
Dining room
Toilet ground floor
Kitchen
Living room
Hall first floor
Toilet first floor
Rest of the house
Bedroom
Bed
In Bed
Out Bed
In bed difference
Out bed difference
Stage A
Stage W
Stage L
Stage D
Stage R
Stress percent
Time in bed
Time light sleep
Time sleeping
Time deep sleep
Resting heart rate
Noise average
Luminosity average
Temperature average
Sleep efficiency