# Mobility Task Description

Data Mining Lecture Winter Term 2019/20

## Motivation

The key target of mobility research is to understand mobility behaviour. One approach is to formulate mathematical models and train these models on observed data. In the past it was common practice to ask people to fill pen and paper travel surveys to collect useful data. Nowadays smartphone assisted solutions help to overcome many disadvantages which pen and paper surveys inherit.
Smartphones are equipped with different sensors (GPS, Accelerometer, Gyroscope, Magnetometer, etc.). During this task you will learn how some of these sensors behave during typical mobility actions. You will collect sensor data with an android smartphone application and analyse the collected data.

## Milestones and Schedule

Milestones 1 and 2 are the initial ones and will give you some time to read the description files and get used to the field. There will be a short Q&A session at the 23rd of Nov.
The results of Milestones 3 and 4 will be presented at the due dates in lecture as a group.

- MS1 due 22.10.2019: Install the App and set your username as **“dm19_lastname”**
- MS2 due 3.11.2019: 1-2 recorded trips per person uploaded. The data will be provided 1-2 days afterwards. You will be contacted via mail and download the data.
- MS3 due 12.11.2019:
    - You prepare first basic analysis for presentation of your trips (only consider the trips from your group for this MS). Prepare one slide per trip and prepare to present these to the class. Don’t put too much effort into this presentation, keep it simple!
    - You have read the task description in detail and prepare questions to be discussed in lecture if necessary. Q&A session!
- MS4 due 15.12.2019: All trips are recorded and uploaded.
- 17.12.2019: Q&A session in class
- MS5 due 05.01.2020: Upload your final report, one pdf file per group.

- MS6 06.01.2020 - 08.01.2020: Peer Review
- MS7 due 09.01.2020: Prepare a slideshow to present your results to us and your colleagues during the lecture. You will have 15 minutes time for your presentation and should split the presentation time among your group members. You submit your presentation slides.

# Overview

To fulfill this task you will form groups of 3-4 students. The task consists of the following subtasks:

1. App Installation
2. Data Collection
3. Data Preprocessing
4. Visualisation and Basic Statistics
5. Clustering/Classification
6. Documentation

The weblink to download all the essential materials:
https://tinyurl.com/yy447kop
Or the long version:
https://drive.google.com/drive/folders/1eduh5euOe2ETz-xVrWvBC8CStRhwdV6M?usp=sharing

The weblink to download the data:
To be defined

We use R (and RStudio) to present the basics of the task, but feel free to use the language of your choice.

# Data Access

After you finished the recording of a trip you will upload it. You will get two data chunks for download, after MS2 and after MS4.
See the file Data_Mining_2019-Get_Started.html for details about the data and how to load, preprocess and work with it.

# Detailed Task Description

The tasks consist of mandatory and optional parts. If there is no information given it is mandatory. The more optional tasks you complete and the more creativity you prove in solving the task the better for your final grading.
The whole task, and all subtasks concentrate only on the **accelerometer** sensor data. Performing Subtasks 1-6 for accelerometer is mandatory. It is optional to use GPS or any other sensor data (if available) to enrich the data set. Most likely this would improve your methods performance, if you add sensor data (especially GPS), however it's optional.

## Subtask 1: App Installation

Install the android application on your private smartphone. If you don't have an android smartphone contact us maximilian.leodolter@ait.ac.at.

Open https://tinyurl.com/yy447kop

Download the APP to your smartphone. Navigate to the APK and install it. Maybe you need to adjust your device settings to install from third party (this should help http://www.aw-el.com/install_android_apps.htm).

When you open the app for the first time you need to accept the authorisations. Set the username as **"dm19_lastname"** (use your lastname, all in lowercase!)**.** Set a password you remember. You're ready to go.

## Subtask 2: Data Collection

Each person collects 1+5 trips. Each recorded trip must fulfil the following criteria:

- Your first trip record is a walk-only trip. Record a trip of 5-10 minutes where you just WALK.
- The remaining 5 trips are multimodal trips. Each should last at least 10 minutes and not more than 60 minutes (optionally you can collect more trips, but they must all be annotated correctly).
- Very important: If you annotated incorrectly then press CANCEL and restart the trip recording!

- Choose the transport modes for your trips out of the following set: (WALK, BICYCLE, CAR, BUS, TRAM, METRO, TRAIN). Ignore the additional buttons for WAIT and UNKNOWN.
- If one of the transport modes of a trip is a public transport mode (metro, tram, bus, train), then this trip must consist of at least two trip stages (so include at least one change), e.g.: WALK - METRO -WALK - BUS - WALK. The walking segment in the middle is the change from metro to bus.
- If your trip is mainly a car or bicycle trip, then record at least one walking segment before (or afterwards) riding the bike/ driving the car.
- The trip should be unique in the set of your trips. It's ok to record one trip in two directions. For example: record from Hütteldorf to Stephansplatz via U4 and U1, and the other way round, from Stephansplatz to Hütteldorf. But don't record Hütteldorf to Stephansplatz by taking the same transport modes twice.
- Don't record your daily routine.
- At least one trip should include a metro or tram going underground (somewhere where you would expect weak GPS availability).
- For clarification, TRAIN includes S-Bahn, Fernzug, CAT, Regionalzug, ÖBB, Westbahn, etc.. Badner Bahn would be TRAM.
- Apart from annotation, you should use your smartphone as usual. Try to not deviate from your habits, e.g. if you normally read the news on your phone when you're in the metro, then also do this during recording.

## Subtask 3: Data Preprocessing

Before you start with the data analysis, some preprocessing steps are necessary. Take a look into the file Data_Mining_2019-Get_Stareted.html to find helpful functions and packages for working with the data. It is **mandatory** to complete all of the following steps:

- The time stamps are typically not exactly equidistant. Deal with this problem by interpolating the observations at equidistant time stamps with a sampling rate of 100Hz.
- Reduce the sampling rate by PAA (piecewise aggregate approximation) to 10Hz. Optionally you could try other values than 10 for the PAA and see how the results change.

- The first and last few seconds of each trip will be affected by the smartphone handling when you start and stop the trip and put the phone into it’s supposed position. That’s why you must discard the first and last 30 seconds of your trips and NOT consider these for further analysis.
- To make the analysis independent of the smartphone’s orientation calculate the 2-norm of the 3-dimensional accelerometer signal for each point of time.
- The timestamp is the number of milliseconds since 1970-01-01. For the plots and statistics make it human readable.

## Subtask 4: Visualisation and Basic Statistics

Every proper data analysis starts with a visualisation of your data. Find your own way to visualize your trips to fully understand it and explain why you chose these kind of plots ( probably you will need more than one type of plot).

**Mandatory:** You should be able to plot the accelerometer sensor data of a selected trip (x, y, z, total).

**Optionally**:

- Plotting time series of moving statistics with sliding windows (e.g. moving average).
- Plot the GPS data of your trips on a map, to actually see where the recorded trip took place.
- Plot sensor data of different sensors (in this case do not forget the preprocessing steps).

Start with basic statistics of your data (mean, median, variance, quantiles, ACF AutoCorrelation Function, etc. ) to get a deeper understanding of your data and of characteristics of different transport modes.

## Subtask 5: Classification

- Split each trip into segments of 10 seconds (after preprocessing steps this should be a 1-dimensional time series of 10 * 10 = 100 observations). Skip segments that are
  a. shorter than 10 seconds (typically the last segment of a trip), or
  b. bimodal (segments covering two transport modes, at the change of the transport mode)

- Build 2 classifiers:
  a. Single Classifier model (**SCM**): Classify the transport mode of the segments based on the accelerometer data (and optionally on all the other data- GPS, etc.).

   b. Ensemble Walk Classifier model (**EWCM**): The same as in (a.) but solve it with an ensemble approach by first training a model to decide whether a trip segment of 10 sec. is walk or non-walk, and then classify all the non-walk segments. For this model report the model performance of Walk-NonWalk model and the complete ensemble model in the final report.
   c. Optional: think of other hierarchy structures of the transport modes and which could be beneficial for a classifier model (e.g.: first classify into rail- road - active, and then continue inside these umbrella-categories)

- Feature extraction:
   a. For feature extraction we recommend the following two approaches that you can use as stand-alone, combined, or propose a completely different solution.
      - Distance approach: classify time series snippets based on distances of the time series to prototypes/ k-nearest neighbours. We name only two of the many possible distance measures: the Euclidean distance, and (most likely performing better) the DTW - Dynamic Time Warping Distance (for questions regarding DTW have a look at this R package and it's documentation that you can use for this task https://cran.r-project.org/web/packages/IncDTW/index.html ).
         - To clarify, for the distance approach your data is a list of vectors, each vector has the dimension T x 1, where T is the length of the n2 accelerometer time series segment after preprocessing, so 200. And the length of the list is the number of time series segments.

      - Feature approach: It is completely up to you what features you extract (also the distance to a prototype time series can be interpreted as a feature). The minimum requirement is to extract 10 features (e.g. mean, stdev, ACF, max, min, percentiles, …) and train a model of your choice. Be creative and think about features that could be distinct for single transport modes (e.g. you will see that the stdev is a valid feature to separate Walk and Bicycle from the other transport modes). The model you train can be e.g. SVM, Decision Tree, Neural Net, Logistic regression ... You can have a look at the R caret package https://topepo.github.io/caret/index.html or https://cran.r-project.org/web/packages/caret/index.html that should offer you a nice user interface to try out different models. No matter what model you choose, you should be ready to explain the model. Do not use a blackbox without understanding what the blackbox is doing! (**Optional**: Perform Feature Selection to improve your classifier's performance.)

- To clarify: for the feature approach, your data is a list of vectors, where each vector has the dimension (K x 1), and K is the number of features you extract, so at least 10. And the length of the list is the number of time series segments.

- Use the data of all students in class. Split the data tripwise:
   a. Use 5-fold cross validation for evaluation. So train the model 5 times on all trip segment data originating from 80% of all trips, and evaluate it on  the data from the remaining 20%. Important: split the data by trips, otherwise you will get overfitting issues because you might try to predict the data snippet of the same trip which is most likely highly correlated! Also, take care of the distributions of the transport modes in the 5 folds, they should be balanced.
   b. Make your work reproducible! Use a random seed when you split your data randomly.  There will be a peer review round where your colleagues will need to reproduce your results, if this is not possible then your total scoring will decrease.
   c. For each classifier: report the following performance measures: accuracy, precision (macro and weighted), recall (macro and weighted), F1-scores (macro and weighted)
   d. Compare the performances of your models
   e. Optional: also describe which of your model attempts (especially those that you don't describe in detail in the final submission) perform worse.

- **Hint:** The more you take care when collecting the data to correctly annotate the records and cancel false annotated trips, the better the classification results will be.

## Subtask 6: Documentation

Finally you describe your group's work and results in ONE pdf document. The document must include:

- The names, Matr. Numbers and Tokens of your group members
- A summary of the trips you collected (number of trips, the amount of time you recorded, selected transport modes)
- The results of Subtask 4
- The results and a detailed description, including code, how you got to the results of subtask 5.

Please keep in mind, **the more optional** tasks you fulfill and **the better your classifier’s** performance, the **better is your grading!**