Reims tram – exploratory data analysis

# Prerequisites

* Install Anaconda (Python 3.x, numpy, scikit-learn, pandas …)
* Glance through all the documents in the data folder, select the important one(s) and read them thoroughly
* Try to really understand physical meaning of all the columns in the data and all the derived features, especially all the powers and interconnections between them
* Create Jupyter notebook with:
  + intense descriptions of everything being done in the comments and in the notebook itself; a rookie should be able to rerun the notebook without any help and reproduce your results
  + most of the code should be done in the form of classes or at least def’s (procedures)
  + export images in PNG (for later usage – the images should be easily exported also in PDF; the code should be easy to run and reproduce results)
* Create Word report:
  + include **Shearer (2000) – CRISP-DM** description and especially part about exploratory data analysis (data understanding)
  + describe the setting of exploratory data analysis (which software is used)
  + include description of raw data
  + include descriptions of all data manipulation
  + include comparison of our work (transformation) with the processed csv file
  + include preliminary results
  + include different factors, learned from data
  + include graphs and tables of all data analysis and results
  + add discussion and in-depth insights about the data
  + properly reference all the figures
* Finally (or by the way): update README.md in the main directory of the repository with the description about how to run exploratory analysis (a document for a rookie to know, how to install the environment and run the experiments)

# Sensor Calibration

Based on formulas from PNG image (Calibration.png). *Zero values* are calculated with average values from first 30s of data. First 60s of data is being dismissed (can contain different internal tests and is therefore not valid).

# Data Transformation

Based on the values from (1) we make the transformed values of raw data – i.e. power calculations. Try to add speed and acceleration based on GPS data. Compare results with processed CSV data. Can we put different columns together and observe if the data is consistent (i.e. if all the sub-powers sum up into total power)

# Basic Visualizations and Basic Values

* time series visualizations (full time series, a few minutes of data)
* discover some errors in the data, display them
* is there any missing data?
* for each time series calculate:
  + min
  + max
  + avg
  + variance
  + noise (based on variance on an interval, where the value doesn’t change much)? how to do this?
  + others?

# Possible Data Cleaning (noise removal for U/I sensors)

Can we apply simple Kalman filtering to the data and remove noise? Can we do some checks? For example – compare all the separate powers with overall power for a particular time stamp (for the raw data and for the filtered data).

# Other Exploratory Analysis

* correlation matrix
* display multiple time series of heavily correlated sensor on the same graph
* scatter chart for correlated and non-correlated timeseries
* typical profiles
* heatmap by hour (?)

# Simple Modelling

Try to model any kind of power with existing data with linear regression. Display weights of different sensors in the model. Try to model for different time horizons – maybe in interval of 10 seconds until 5 minutes ahead.