**Good morning and welcome everybody**

My name is Pieter-Jan and I study the master artificial intelligence. Previously I studied statistics and economics at Ghent University. So, I am going to speak a little bit about the progress of my thesis which is about load forecasting. It a in cooperation with Emeryville, which is a company that performs research to sustainable energy and intelligent systems.

**Table of contents**

So what’s on the menu for today, first I am going to introduce the problem and motivate the thesis and also describe the aims of the thesis. Then I will focus on what’s been done in the literature.

Next, I will go to the actual data analysis part and finally I will end with some discussion of my planning for the upcoming weeks/months

**What is the motivation?**

Load forecasting has been a fundamental issue in the planning and operation of power systems. Having accurate point forecast places a company at competitive advantage. Though, in this competitive environment, point forecast are not enough anymore. Often in many operational and planning tasks, probabilistic forecasts are needed. This is also reflected in the literature, focus in the past has mainly been on point forecasting, but in recent years research on probabilistic forecasting has taken off rapidly.

**The goal of the thesis** is in a first stage to analyze the data in terms of quality and quantity. Secondly, perform some data exploration and try to visualize interesting behavior. Finally, prediction: first try to develop a point forecasting models and then also a model that generates probabilistic forecasts.

**Literature review**

For the literature review I focused largely on the global energy forecasting. There have been 3 of them: on 2012, 2014 and 2017. One of the goals of these competitions was to gap the bridge between academic research and industry practices. Participants were asked to document their methodology in a small report. An the reports of the best performing are published. I focused on the first two competitions, 2012 was on point forecasting and 2014 was on probabilistic forecasting. I have summarized for each of these competitions the techniques used by the top 3 performing teams in three different tables: One tables focuses on feature engineering and important features, the second focuses on point forecasting and the last one focuses on probabilistic forecasting.

I found these competitions a fun way to explore the field of load forecasting and get familiar with the best practices used in the industry.

**Data Cleaning**

So let’s now move on to the more fun part.

I obtained data from Fifthplay and Fluvius . The Fifthplay dataset are the data measured by Energyville. The dataset from Fluvius is used as sort of control set to validate the data of Fifthplay. So for Fifthplay, I obtained PV generation, injection and consumption for transformer 1 and 2. The same for Fluvius, but data was already aggregated for the transformers.

The total consumption of the building was then calculated as follow: PV generation minus injection + consumption. The resolution of the data was 15 minutes, so 96 observations equals one day. The data started as early as end of may 2017 till the beginning of 2020.

In general the data was pretty clean, there were some minor problems such as a gap of almost one month in the beginning of 2018. Then I used the data from Fluvius to impute. The figure here shows the consumption of Fluvius and Fithplay and the difference between the two. The difference is not exactly zero and this is expected because the measurement systems of fluvius and fifthplay are not on the same place.

**Exploratory Data Analysis**

Exploratory data analysis, the first thing I did was splitting data in a training, validation and test set (60 %, 20%, 20%). The splits were done chronologically to preserve the temporal structure in the data. The EDA was performed on the training data.

The top figure show the consumption over time and the figure below shows the distribution of the consumption. You might see a small increasing trend, certainly in the beginning, but then it stabilizes. Also the distribution is positively skewed, but that’s mainly because of consumption between day and night and week and weekend and this is also illustrated on the next figure

Then I looked at some profile plots. So the figure on top shows the mean consumption and the red shaded area is the 1 standard deviation. The figure in the upper lower corner shows the mean consumption throughout different seasons. So the highest consumption is during Winter and the lowest consumption is during Summer which seems reasonable. Then the plot on right is zoomed in version of the plot on top.