What real-world problem do you want to solve?

**Problem framing:**

By scrutinizing the dataset and exploring all variables of electrical consumption in a very household within nearly four years, we have detected some trends, seasonal features and other noteworthy details that will allow us to predict future energy consumption demand with some degree of accuracy.

We can predict for the active energy consumption for the nextday on an hourly basis, next week on a daily basis, next month on a daily basis, next quarter and even next year on a monthly basis.

What data mining task(s) would you model it as, association rule mining, clustering, and/or classification?

**Modeling Overview:**

**Types of modeling:**

For time series forecasting, there are classic statistical methods such as:

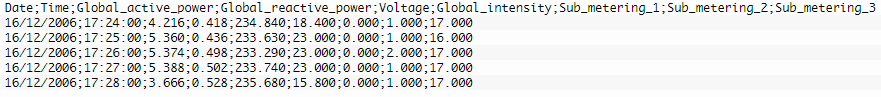
* **Autoregressive methods:**  
  The ARIMA (Auto Regressive Integrated Moving Average), this model could be extended to most time series data types. For one or more autoregressive (AR) terms (its own lagged values) and/or one or more moving average (MA) terms, ARIMA can be seen as a multiple  regression model.
* Using regression, we can plot graphs and predict the future demand of energy consumption.

We can reframe our time series problem as a supervised learning problem as  well as unsupervised machine learning problem.

* **Supervised learning problem:**  
  Predicting the time series can be interpreted as a regression problem in which  the output variable ("active power") is a real value.
* **Unsupervised learning problem:**  
  When we have unlabeled data from time series, we can use many unsupervised algorithms, such as: k-nearest neighbors, for prediction.
* **Deep learning problem:**For time series approximating a mapping function from input variables to output variables is important.
* Methods such as CNN (Convolution Neural Networks) and RNN (Recurrent  Neural Networks) offer great promise for forecasting time series.

What data set do you plan to use? Provide a sample data table to include instances and attributes. Provide important data summary that is relevant to your problem modeling

Below we present the first five rows from the raw data file.  
The data consists of 2075259 observations with 9 variables representing different energy measurements.

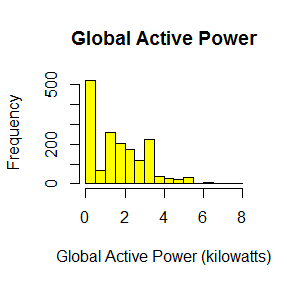


The data does have missing values about **1.25%** of the rows (about 80 days), with some dates having some missing values.  
We decided to replace these missing values with the median of the actual column values.

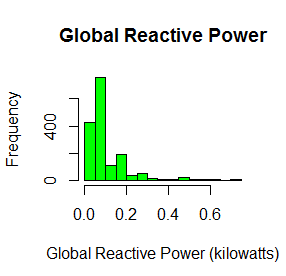
A good way to understand the data is to analyze it so as to figure out some consistent patterns or trends and to understand whether seasonality is relevant or whether there is evidence of  certain cycles.

Below are few of the histograms that we have plotted:

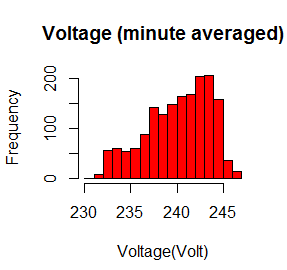
1) Global active power



2.) Global reactive power

.

3.) Voltage



How would you evaluate your data mining result?

We provided the initial visualizations of the household power consumption dataset, which  explains energy use over four years for a single house. The task at hand is to estimate future  energy usage requirements.

Knowing the amounts of energy usage is extremely important for both customers and energy  suppliers. In fact, it is very useful for consumers to know their consumption to better understand and  manage their energy demands.

Also, analysing the data of energy consumption is could help a utility company to predict probable future demand and apply the costs for electrical units on the market of electrical energy.

Draw an illustration that summarizes your project proposal and submit it with your written project proposal.

Better the Model

***KARTHEEK SUNKARA***

***PRANKUR GARG***

***SNEHAL GHATATPANDE***