# **Data Science Project: Gas Sensor Data for timely Prediction**

**Introduction**

This document is an outline of the “Gas Sensor” data science project to familiarize students or employees with sensor data and practice building a prediction algorithm.

**Context of data collection**

A group of researchers placed 10 sensors (8 MOX sensors, 1 temperature sensor and 1 humidity sensor) in one of the researchers’ residence, recording the sensor levels for 2 years, trying to determine if the MOX sensors would detect the presence of odors in a room. The data has the recordings of 3 situations:

1. measurements of sensors before, during, and after the presence of a **banana**
2. measurements of sensors before, during, and after the presence of **wine**
3. measurements of sensors with **no odors** in the room (baseline)

For more information and for the dataset, here are important links:

Relevant paper:

<https://arxiv.org/pdf/1608.01719v1.pdf>

Dataset and description: <https://archive.ics.uci.edu/ml/datasets/Gas+sensors+for+home+activity+monitoring>

**Objectives**

Using the dataset available only, determine the source of the odor (banana, wine or background) as quickly as possible. Here are some steps to help develop this prediction model:

* Understand and visualize the data
* Identify relevant data and features
* What will be the mechanics to detect the right odor as fast as possible?
  + Question: Do you need to have all of the data to identify the odor?

Answering the following questions will help determine the structure of the model:

* + **What are the in-scope data?**
    - In the process implemented in python code, we are not going to use t0 (the point when induction started), dt (interval that this induction lasted) and Date (Day, Month and Year when this induction was recorded) from the metadata. Other than that, we are going to use all the data one or other experiment.
  + **What are the input variables/features?**
    - In these experiments, we are going to use readings of 8 MOX sensors and 1 temperature and humidity sensors each. These variables will be features for our experiments.
  + **Which input variables are most helpful in predicting the type of odor? Alternatively, which input variables help predict the type of odor the fastest?**
    - We have used correlation to find the most important features among all 10 features. After reducing least important features, we worked with the variables from 5 MOX sensors and 1 temperature and humidity sensor each. This made total number of features/variables to 7 from 10. These least relevant features for our pupose not only make processing time higher, but also reduce efficiency of the system.
  + **What kind of predictive models should you use? Which one performs best?**
    - In these experiments, intially, we have used SVM with different size of features and after that experimented with Random Forest. In our problem, Random Forest performed best.
  + **Should this model be implemented in real time, how long would it take for your model to correctly determine which odor is present?**
    - I believe, we can implement this model in real time, however, we need to first define interval of thee consecutive readings of the sensors and then need to experiment with different size of data from the time of induction and test with them. We might need to optimize the system for the speed vs efficiency.