# Data Analytics Task Description

On Monday 16th March 2020, each team and its members will be required to deliver a presentation on the analysis of the dataset referred to in this document. Each team will have seven minutes to present their work. This presentation will be followed by questions asked by the panel to assess the team’s work. This presentation will weigh towards 10% of the overall module grade (group mark).

In the first phase of this project, you designed and built an end-to-end system to emulate an HVAC system. In this second phase of the project, you are required to understand the behaviour of such a system and augment it with predictive control through the use of data analytics and machine learning, towards the ultimate goal of autonomous operation.

## Dataset description

For the purpose of this second phase of the project, you are provided with three datasets, which are available on Moodle, namely Raw Data\_Room 101, Raw Data\_Room 103, and Raw Data\_Room 144. These three datasets (spreadsheets) contain the recordings from individual sensors that were deployed in an office building in central Athens, in Greece (courtesy of one of our industrial partners). For the purpose of this project, you may assume that the building is located anywhere within central Athens. The position and orientation of the different rooms within this building is illustrated in Figure 1.

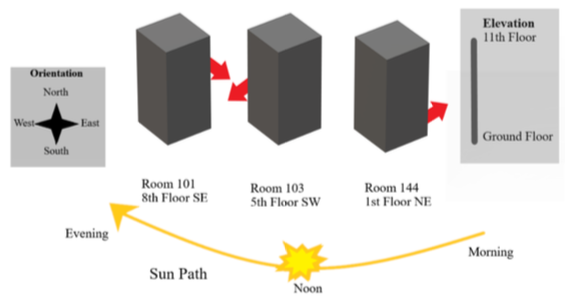


Figure 1: Position and orientation of rooms in designated office building

Each of the three datasets has the following seven features:

* **RoomTemp**: Room temperature measured in °C, with 0.5°C precision.
* **SetTemp**: Setpoint temperature measured in °C, with 0.5°C precision. The setpoint temperature is essentially the “thermostat temperature” which can be varied by the user (but can also automatically return to preconfigured values in the HVAC system).
* **OnOffState**: Indicates whether the HVAC system is on or off, with one (1) representing On and zero (0) representing Off.
* **OperationModeState**: Operation mode of the HVAC system, which can take on one of four values; one (1) for Cooling, two (2) for Heating, three (3) for Ventilation and four (4) for automatic mode.
* **OUTAIRHUMD**: Outdoor Air Humidity measured in % Relative Humidity.
* **OUTAIRTEMP**: Outdoor Air Temperature measured in °C, with 0.005°C precision.
* **AlarmSignal**: Returns a value of one (1) if there is a problem with the HVAC system, otherwise is zero (0).

## Aims and objectives

This phase of the project aims to answer two questions; First, can we identify and extract unique and hidden patterns and trends from the given datasets? Second, can we predict future events with a high level of accuracy and interpretability?

The following guiding questions may help you during this data analysis journey:

* Can you identify any patterns and/or trends in the dataset that would generate unique and informative insights? Can you create an inference model explaining the underlying relationship between the different features? In other words, what is the impact of varying the value of one feature on the value of another feature?
* Can you create a prediction model that would feed into your end-to-end system in order to provide accurate and autonomous control and actuation?
* Can you import external data to enhance the existing datasets, for example additional weather data or other event data (with reference to the specific dates and times indicated in the provided datasets)?

From a user or business perspective, the features in the given datasets can inform two metrics; thermal comfort and energy efficiency (without requiring meter readings):

* The ASHRAE (American) and CIBSE (British) building standards define ranges and threshold levels for different environmental variables that affect user thermal comfort. Research has also shown that variations in setpoint temperature are strongly correlated with users’ behaviour (in terms of comfort).
* The same standards also define ranges and threshold levels that determine optimum energy efficiency in a room or building, even without access to energy meter readings.

**Objective**: Your objective is to determine the thermal comfort and energy efficiency levels of the rooms in the designated building, as well as the ability of the HVAC system to adjust and maintain these levels. To this end, you might wish to take the following points into account:

* How does the room temperature compare to the setpoint temperature? Are the trends similar or different for the given rooms? Does seasonality affect the trends in any way?
* Given the current state of the room’s architecture and HVAC system’s functionality, can we predict future values of room temperature and setpoint temperature accurately? Is this prediction accuracy independent of seasonal trends?
* Can we predict the ideal setpoint temperature for each room for optimum thermal comfort and energy efficiency based on building standards?
* Can we incorporate additional features into our models to improve the prediction accuracy and/or interpretability of our models?
* Can we generate plain-text reason codes to automate the extraction and interpretation of hidden patterns and trends across the entire dataset?

You are expected to use an industry-standard tool for the purpose of carrying out data analytics and machine learning. Such tools may include MATLAB, R or Python but not Microsoft Excel. You may also use IBM’s Watson Studio platform should you wish.