**Problem Statement**

Air Quality is commonly measured by means of an index based on the concentration of several pollutants. The concentration of the pollutants, in turn, may be correlated with each other, along with other factors such as temperature, pressure, and wind.

The aim is to detect anomalies in the pollutant concentrations and a few other variables they may depend on.

Note: I initially wanted to analyze the air quality index as well but discarded the idea because it is a known deterministic function of the pollutants’ concentrations

The data is sourced from the ‘Beijing Air Quality Data’ dataset from the [UCI repository](https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data).

**Notebooks**

Two notebooks have been provided on the [Github](https://github.com/MamoonHaqqi/air_quality_data_anomaly_detection) repository:

1. **EDA and Data Prep** – Contains the preliminary data analysis, and the testing of the data preparation sequence and anomaly detection techniques. Not very organized, and is included in case the evaluator wishes to examine the underlying thought process
2. **Solution –** Contains the cleaned and consolidated code, which has two blocks: one for data preparation, and the other for the anomaly detection sequence.

**Note:** Pleasedo not spend too much time on the data preparation block. I intended to do a more thorough analysis including the Air Quality Index and the covariates but could not. Thus, the variables being created are not of critical import

The Beijing Air Quality dataset contains hour-wise readings of the sensors from 2014 to 2017. It has missing data for some of the columns, but no time steps are skipped.

Air Quality Index calculations are typically done daily. As such, it is useful to look at both the hourly and daily data. The **Solution** notebook creates two datasets:

1. df\_c – contains the hourly data for sensors, as well as anomaly detection flags at hour level
2. c\_data – contains the sensor and covariate data averaged daily. Also contains the anomalies detected at the day level

**Notes**

**Anomalies are commonly defined as observations which deviate from the expected or some standard.** I’ve taken some time to put my perspective here to reach a more rigorous definition, to better outline my own approach.

**An anomaly is a data point which does not conform to patterns commonly seen in the data** or **cannot be explained using patterns learned from other data points**. In other words, it is a data point for which a reasonable explanation cannot be derived from the existing data; it must come from some external source to which we may or may not have access.

Two things further color our perception of what is an anomaly or not:

1. The limitations of what algorithms can do
2. Our own ability to identify certain patterns which the algorithms may not be able to pick up

Owing to this, what is regarded as an anomaly would differ depending on the kind of explanations we/our algorithms have to offer.

Below are the mechanisms I intend to explore as a means of detecting anomalies in the current assignment:

1. **Thresholds:** Some domains allow us to define fixed thresholds for metrics, breaching which, the data may be considered anomalous. While not the most flexible, it is the easiest to understand and interpret
2. **Hypotheses/Known Facts:** We may have access to some facts prior to seeing the data, such as the dependence of two variables, or the direction of their relationship. Checks can be put in place to test the data against these facts or hypotheses
3. **Density-based (Outlier Detection):** Using the distribution of the data, we can identify whether newly occurring data points ‘belong’ to the data distribution or not. This is challenging in two cases: a) Where we do not possess enough data or priors to describe the data density in detail, and b) Where the data is high-dimensional
4. **Changes/Shifts in the Data Generation Processes:** Transition points beyond which the data shifts temporarily or permanently to a ‘new normal’. Such occurrences, if frequent, can throw most methods off
5. **Subtleties/Undercurrents:** To understand this, imagine a face recognition system to which people submit their selfies and get registered. If, during two hours of operation, the system is fed deepfakes, it is a subtle (but great) change which may be well beyond the capabilities of that system to detect. In this case, a huge underlying change will have taken place, but the system/its owners may not even notice it. These anomalies are the most difficult to detect because they do not have a defined form, and often we do not know what we are looking for