# **Problem Statement:**

Ground-level ozone (O3), the key ingredients of smog, depends on a sophisticated chemical and physical process as a function of many known and unknown factors. While there are traditional methods used by air quality control scientists to solve this problem, those are not very accurate. Objective of this challenge is to build an **ozone alert forecasting system** using **supervised machine learning algorithms**. This will be useful in issuing warnings to the public before the ozone reaches a dangerous level. You are given two datasets, capturing ground ozone level flags, along with measurement on potential environmental indicators. You can leverage both the datasets for your analysis, as you deem appropriate.

* + Eight\_Hour\_Peak\_Set.csv: Data created based on peak values from last eight-hour interval
  + One\_Hour\_Peak\_Set.csv: Data created based on peak values from last one-hour interval
  + Column\_Descriptions.xlsx: It provides explanation of various columns of the dataset

Datasets contain few missing data points. Please make necessary assumptions and perform missing value treatments, as needed.

You can use a maximum of 75% of the data (with unbiased sampling) for model building. Remaining data need to be kept aside for model validation. Your model will be assessed based on accurate prediction on both model building data and validation data.

Below mentioned links provide required subject matter understanding. Feel free to do additional research on domain knowledge as necessary

* + Wind Components: <http://colaweb.gmu.edu/dev/clim301/lectures/wind/wind-uv>
  + K Index: <http://w1.weather.gov/glossary/index.php?letter=k>
  + T-Totals: <http://www.theweatherprediction.com/habyhints/302/>
  + hpa (Hecto Pascal): <https://www.convertunits.com/info/hPa>

**Solution Evaluation Criteria:**

1. Recall (the proportion of actual Ozone Days which are correctly predicted) – Higher the better
2. Reasoning behind algorithm selection, feature selection, feature engineering and data treatments
3. Notebook legibility
4. Originality of the solution

**Deliverables:**

1. Notebook in native and html format.
2. A report describing steps of implementations, reasoning behind algorithm selection, feature selection, feature engineering and data treatments.
3. Examples of different scenarios (user input).
4. Presentation will be asked if selected for final round.