# Indian Institute of Science


## Scalable Systems for Data Science

Project Proposal

# Analytics on streaming data from Environmental Sensors

Shibsankar Das (ME Microelectronics)
Rajrup Ghosh (MTech CDS)

## ✓ Problem Statement :

*Study of Pollution level from Environmental sensor data in real time*

We will perform real time analysis on air quality, sound, temperature, dust and humidity data to observe the pollution level (Air and Noise) and change in other environmental parameters. Then we can compare the pollution levels of different cities.

## ✓ Dataset:

We will work on dataset provided by Data Canvas [1]

Type of Sensor data:

1. Air Quality
2. Dust
3. Light
4. Sound
5. Temperature
6. Humidity

Temperature and Humidity helps us to determine weather changes for a particular period of time.

Air Quality, Dust and Sound data helps us to develop inference on Air and Noise pollution levels.

## ✓ Importance:

Pollution is one of the critical concerns of modern era. Study of different parameters leading to pollution might help us detect sources and help us make arrangements to cut down pollution levels. Moreover measuring changes in environmental parameters helps in weather forecasts. This requires real time processing of data from appropriate sensors.

## ✓ Methods and Platforms:

Real time analytics to measure mean, standard deviation and other statistics of sensor value will be performed in Apache Storm [2]. As the data set is available for 1 month we can divide the data into history and future. Using the historical sensor data we can forecast the values for future using data analysis techniques and match actual values with predicted results. For visualization of data in real time we will be using Rickshaw [3], which can give us live change of air quality parameters in day and week basis. Real time monitoring of data is a significant aspect in such a scenario.

This kind of analysis of data requires platforms which offers low latency for processing of events. Frameworks like Hadoop and Spark are not suitable for real time event processing.

## ✓ Evaluation:

We will be observing the variation of air quality, noise levels, temperature and humidity for the data available. Pollution control boards of different country gives permissible thresholds for such parameters. We will see how much less or critical are parameters from the allowed standards. We will evaluate our prediction model

for forecasting parameters by matching predicted values with real ones and will be included in the visualization framework.

✓ **Deliverables:**

- Comparison of pollution level and other weather parameters between different cities.
- If air pollution or noise levels are high for certain parts of a city, we will try to find out reasons from the geographical topology of the region.
- Visualize real time data streams from sensors and compare real values with the predicted values obtained from the model.

**REFERENCES:**

[1] Data Canvas, http://map.datacanvas.org/
[2] Apace Storm, http://storm.apache.org/
[3] Rickshaw: A Javascript toolkit for creating interactive time series graphs, http://code.shutterstock.com/rickshaw/