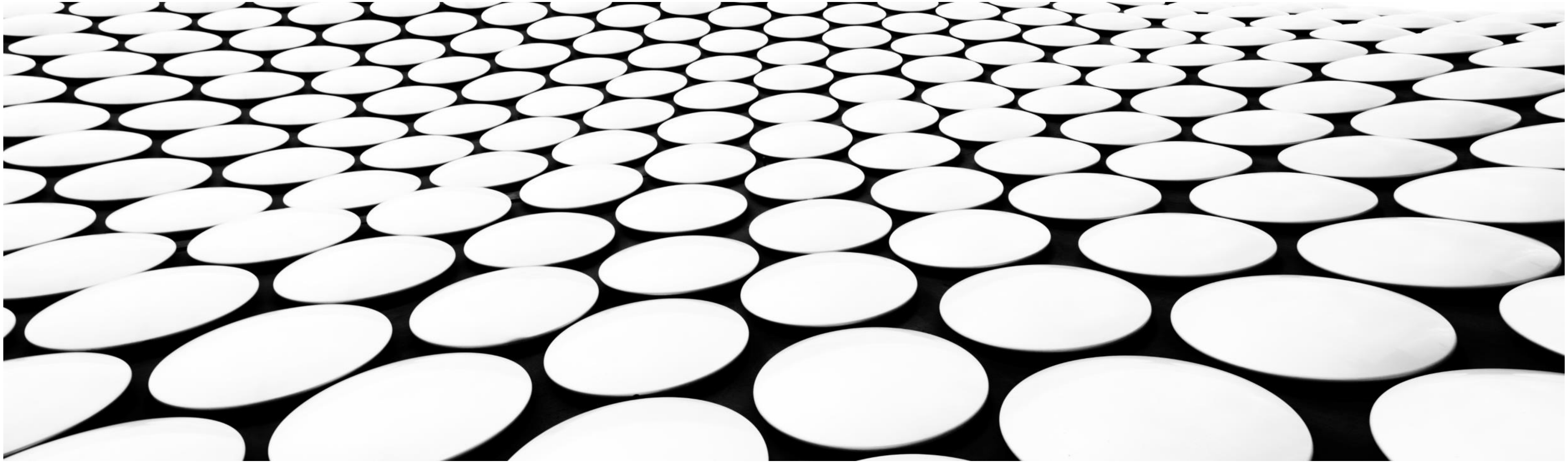
MILESTONE 1 IMPORTING DATASET FROM OPENAQ AND ECMWF

Content
Completion of Milestone 1
Plan for Milestone 2
Ideas for Milestone 2

PARTICIPANT: GORDON RATES

ECMWF **MENTORS: JOHANNES FLEMMING AND MIHA RAZINGER**



MILESTONE 1

Task 1: Get Access to OpenAQ dataset

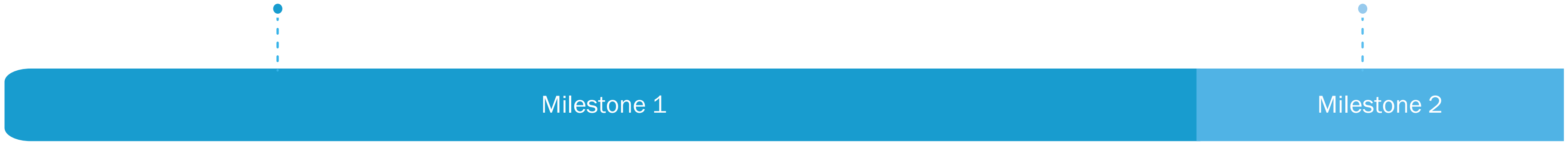
Task 2: Download some data to Local PC

Task3: Construct Presentation

Task 4: Find insights on categories of stations for Milestone 2

Task 1 Statistical Analysis

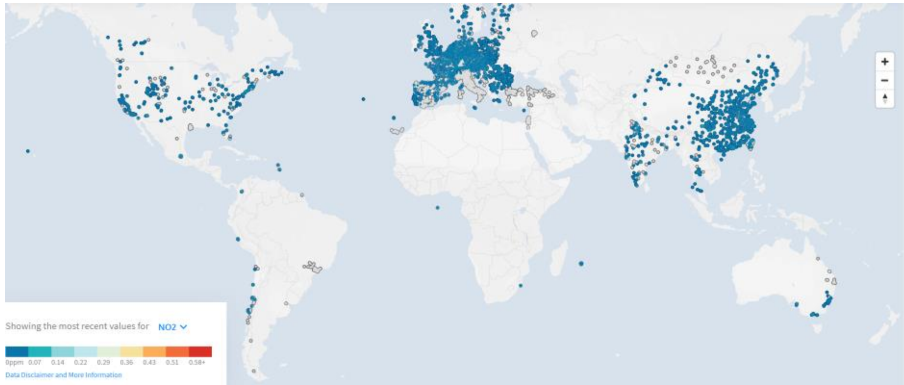
Task 2 Cluster Analysis Methods



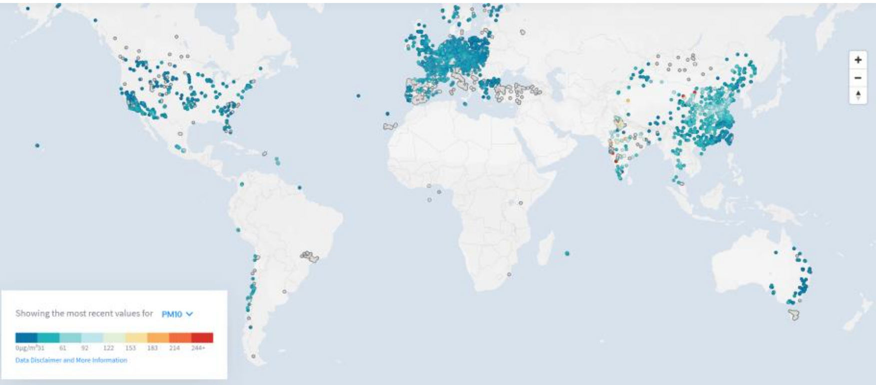
LAST MEETING ACTIONS

- **Feedback to Gordon's presentation and the project plan:**
- Using additional information (emissions, CAMS model results, population density) is very welcome and the mentors Miha and Johannes can help with data access
- Linking the result to the openAQ community developments is to be encouraged.
- Being aware of the large variability in the openAQ data (data coverage period, spatial density, quality control, sensor quality) is important and "lower data quality" in certain areas should not lead automatically to dismissal if no alternative data are available for the area (e.g. Africa)
- **Agreed actions items:**
- Gordon to a send the milestones/deliverables list with an added deliverable about setting up data acquisition in the next couple of days
- Gordon to acquire the full set of openAQ data as soon as possible to identify any data access problems quickly
- Miha will set a cloud machine
- Johannes will set up the next meeting in two weeks time
- Johannes will fill in the form about project description

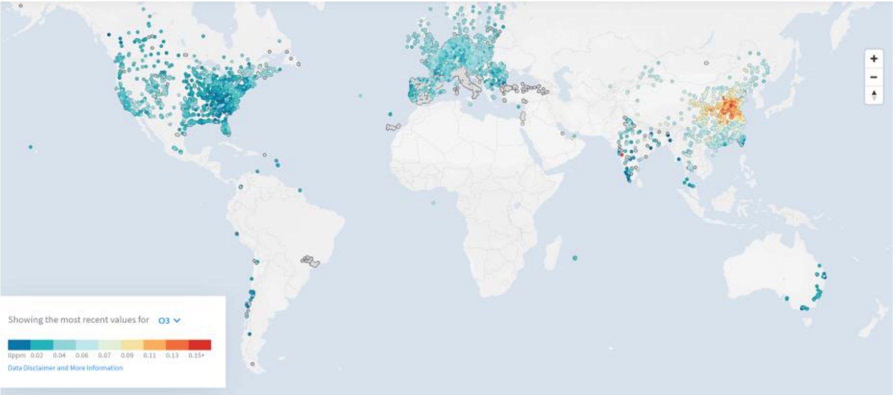
COVERAGE OF OPENAQ DATASET


Showing the most recent values for NO2
0ppm 0.07 0.14 0.22 0.29 0.36 0.43 0.51 0.58+
Data Disclaimer and More Information

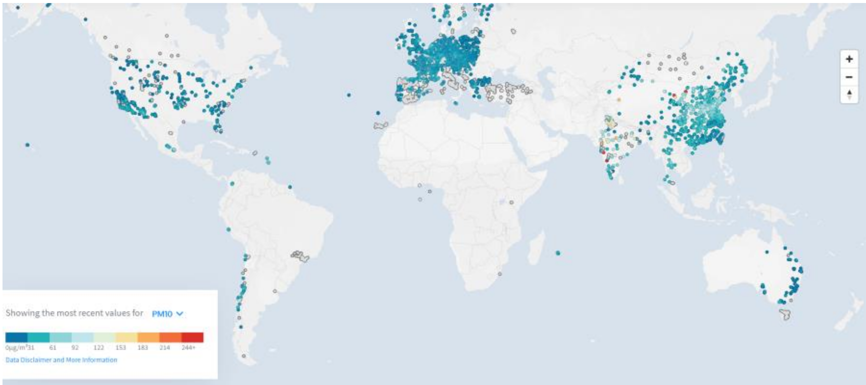
NO2


Showing the most recent values for PM10
0µg/m³31 61 92 122 153 183 214 244+
Data Disclaimer and More Information

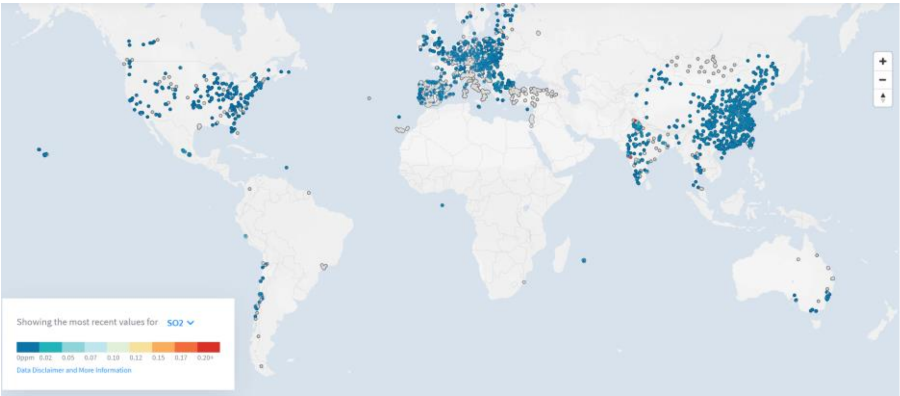
PM10


Showing the most recent values for O3
0ppm 0.02 0.04 0.06 0.07 0.09 0.11 0.13 0.15+
Data Disclaimer and More Information

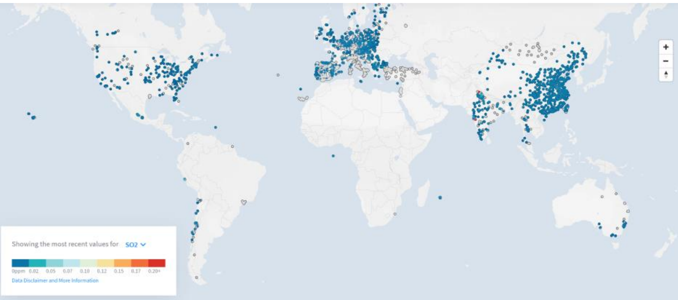
Ozone (O3)

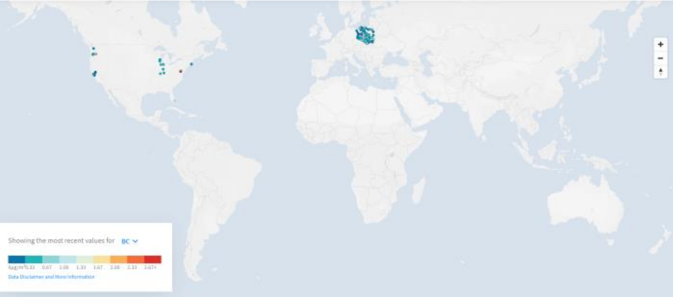

Showing the most recent values for PM10
0µg/m³31 61 92 122 153 183 214 244+
Data Disclaimer and More Information

PM2.5


Showing the most recent values for SO2
0ppm 0.02 0.05 0.07 0.10 0.12 0.15 0.17 0.20+
Data Disclaimer and More Information

SO2


Showing the most recent values for SO2
0ppm 0.02 0.05 0.07 0.10 0.12 0.15 0.17 0.20+
Data Disclaimer and More Information


Showing the most recent values for BC
0µg/m³0.33 0.67 1.00 1.33 1.67 2.00 2.33 2.67+
Data Disclaimer and More Information

CO ↑

BC ↑

POTENTIAL CATEGORIES OF RELIABLE STATIONS

Why More Reliable - Criteria

- Other sensors to validate
- Regulations on Air Quality dataset
- National Networks that are audited
- E.g. European and EU
- e.g. Belgium (use for validation)

Why Less Reliable - Selection

- India (Stations 456 Total Measurements 32,033,248)

NO2 302 Stations, PM10 267 Stations, PM2.5 309 Stations, O3 280 Stations

SO2 291 Stations, BC 0 Stations, CO 302 Stations

- Turkey (Stations 172 Total Measurements 5,640,285 Latest upload 2018/10/13)

NO2 137 Stations, PM10 158 Stations, PM2.5 0 Stations, O3 82 Stations

SO2 142 Stations, BC 0 Stations, CO 75 Stations

- Eastern Europe
- China (Stations)

Uncertainties

Sensor Errors

- Missing measurements
- Faulty Sensors
- Not Calibrated Sensors
- Old sensor
- Sensor damaged
- Sensor covered or hindered
- Sensor not calibrated
- Sensor affect by seasonal difference
- Sensor affect by traffic volume or emission volume
- Sensor affected by humidity

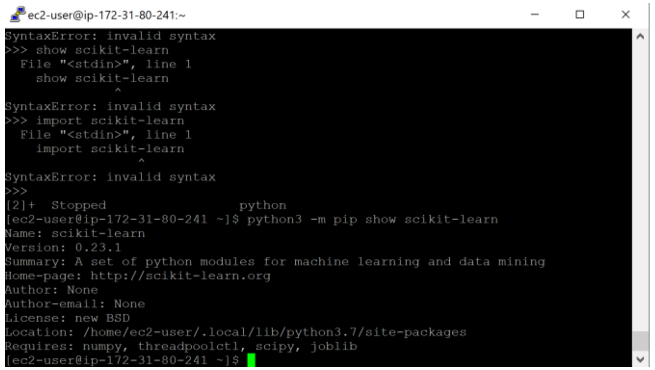
Authority errors

- Authority Fabricated
- Unqualified sensor manager
- Error in data processing or publishing

INSTALLED PACKAGES FOR ANALYSIS

Output

1. Setup AirNode_AQCA folder on EWC
2. Setup FTP and SSH to EWC
3. Copied all of OpenAQ to EWC AirNode_AQCA Folder
4. Access to EWC from Putty
5. Installed Anaconda for ML, Analysis python packages
6. Reviewed scope of available OpenAQ datasets



```
SyntaxError: invalid syntax
>>> show scikit-learn
  File "<stdin>", line 1
    show scikit-learn
              ^
SyntaxError: invalid syntax
>>> import scikit-learn
  File "<stdin>", line 1
    import scikit-learn
                 ^
SyntaxError: invalid syntax
>>>
[2]+  Stopped                 python
[ec2-user@ip-172-31-80-241 ~]$ python3 -m pip show scikit-learn
Name: scikit-learn
Version: 0.23.1
Summary: A set of python modules for machine learning and data mining
Home-page: http://scikit-learn.org
Author: None
Author-email: None
License: new BSD
Location: /home/ec2-user/.local/lib/python3.7/site-packages
Requires: numpy, threadpoolctl, scipy, joblib
[ec2-user@ip-172-31-80-241 ~]$
```

Installed scikit-learn on EWC

Used pip3 install -U scikit-learn

python3 -m pip show scikit-learn

Written a Python program to access S3 and analyse dataset

Installed Anaconda

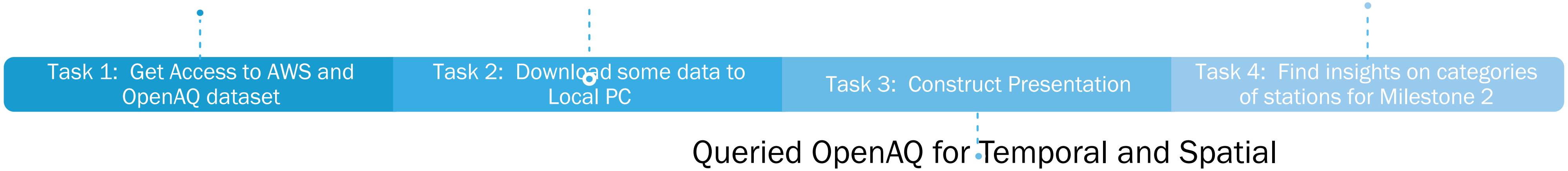
Copied openAD Data to EWC AirNode Folder

Use boto3 to access S3 through Putty on EC2

Used AWS Athena to Query openAQ

Setup an FTP Client using SSH to EWC instance

Reviewed Queries

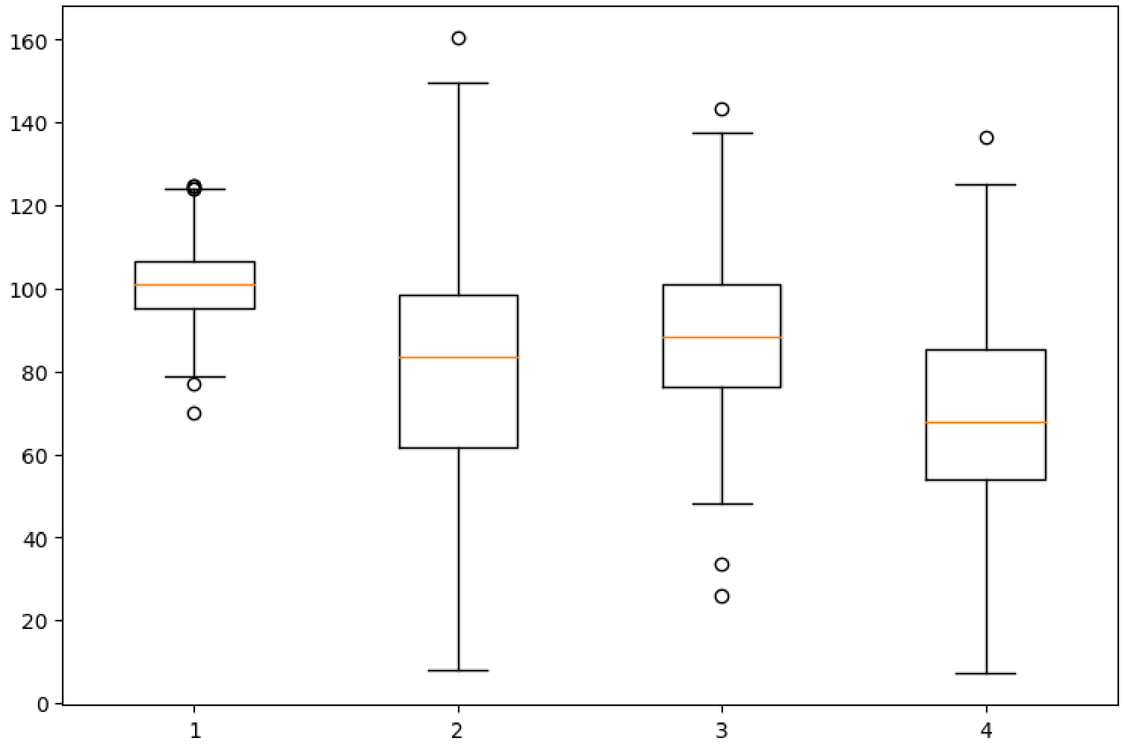


Task 1: Get Access to AWS and OpenAQ dataset
Task 2: Download some data to Local PC
Task 3: Construct Presentation
Task 4: Find insights on categories of stations for Milestone 2

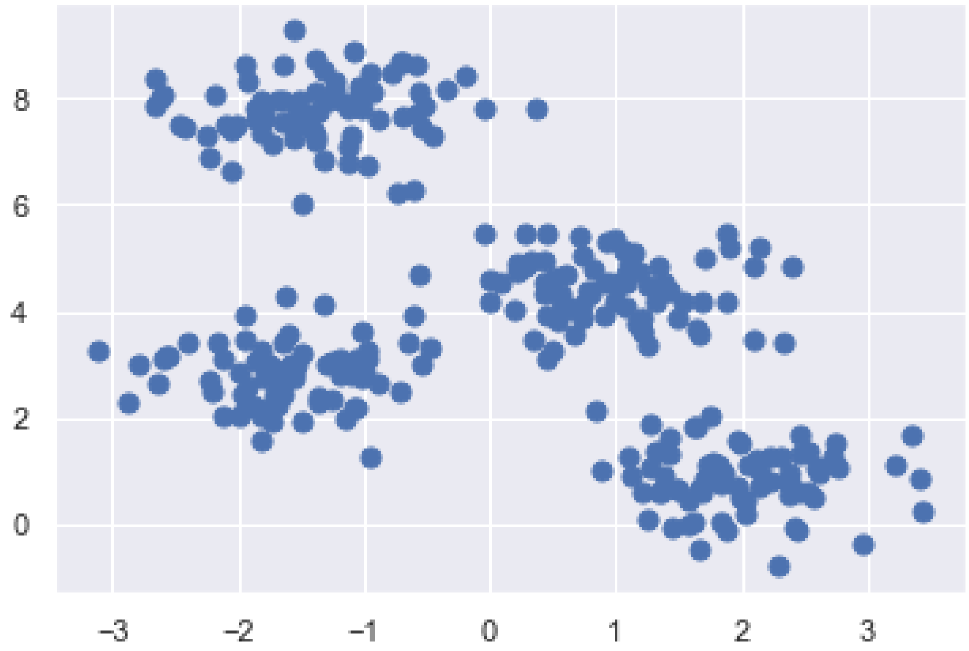
Queried OpenAQ for Temporal and Spatial differences

COMPLETED ELEMENTS

BOXPLOTS



KMEANS



COMPLETED CLUSTER ANALYSIS

Distance-based and Density Methods

Distance Metric : “An appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm. This aspect of the problem … depends on domain specific knowledge and is less amenable to general research.” Hastie, Tibshirani and Friedman's Elements of Statistical Learning

Python Application available:

1 Affinity propagation clustering

2 Agglomerative clustering

3 Kmeans

4 Mini Batch Kmeans

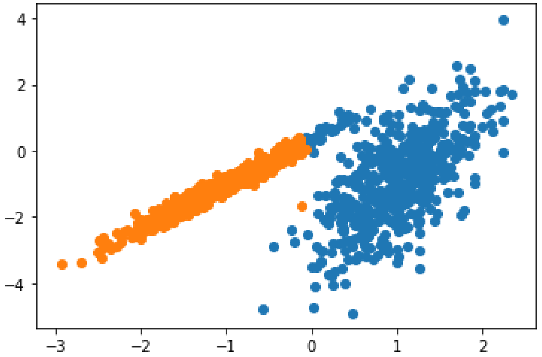
5 Birch

6 Mean Shift

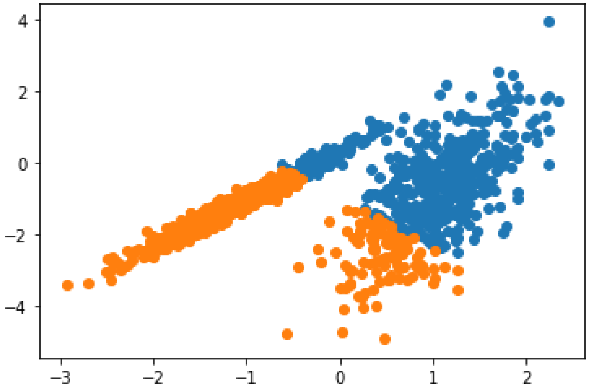
7 Gaussian mixture

8 Density-based spatial clustering of applications with noise (DBSCAN)

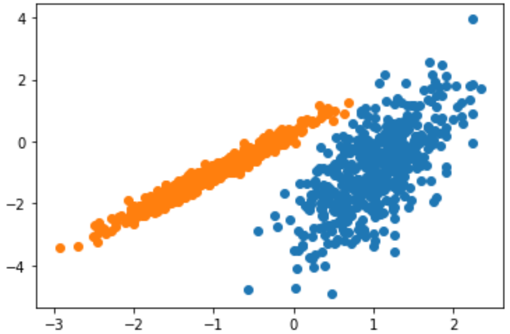
Same Dataset



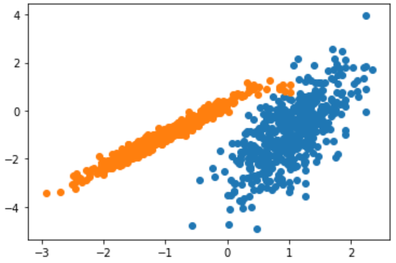
Spectral



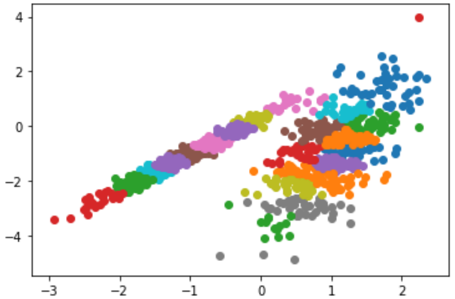
Kmeans



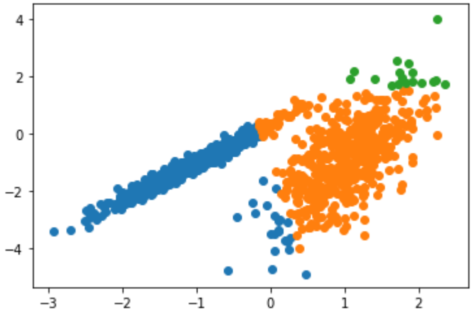
Gaussian Mixture



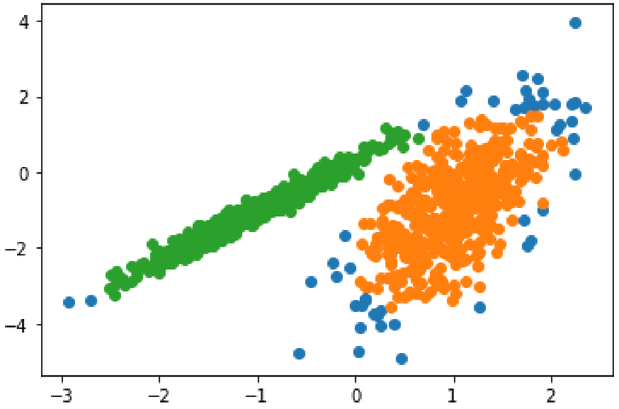
Birch



Affinity propagation

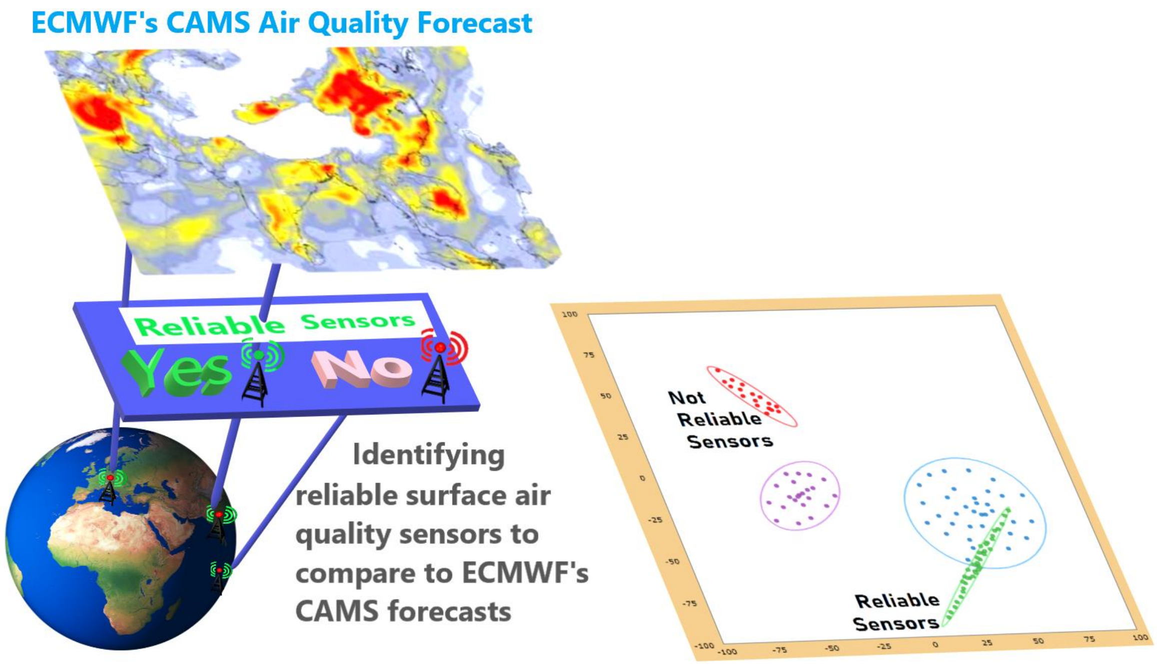


Mean Shift



DBSCAN

PARTICIPANTS FORM



Identifying reliable surface air quality sensors to compare to ECMWF's CAMS forecasts

Project Description

Validating and removing errors outliers from surface air quality observations from individual sensors so that these observation can be compared to ECMWF's CAMS air quality forecasts.

By clustering analysis on these observations more reliable observations can be identified. Enhancing these observations by attaching data about factors that affect air quality these observations can have more credibility about their accuracy.

CAMS lacks credible surface air quality observations in many parts of the world, often in the most polluted area such as in India or Africa. Some observations are available for these areas from data harvesting efforts such as openAQ but there is no quality control applied to the data, and it is often not well known if the observations are made in a rural, urban or heavily polluted local environment.

This information on the environment is important because the very locally influenced measurements are mostly not representative for the horizontal scale (40 km) of the CAMS forecasts and should therefore not be used for the evaluation of the CAMS model.

PLAN TO CATEGORIES STATION IN MILESTONE 2

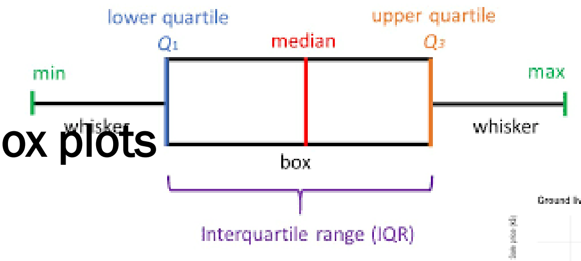
IMPORT Datasets -> Choose Variables - > Identify Clusters -> Obvious Outliers (Global Outlier)

Step 1

Step 2

Visual analytics

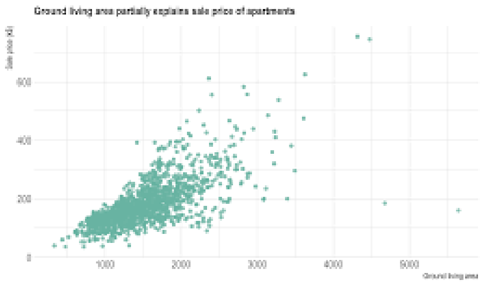
Cluster Analysis (K-means, Expectation Maximization, Canopy, Farthest First and Hierarchical clustering)

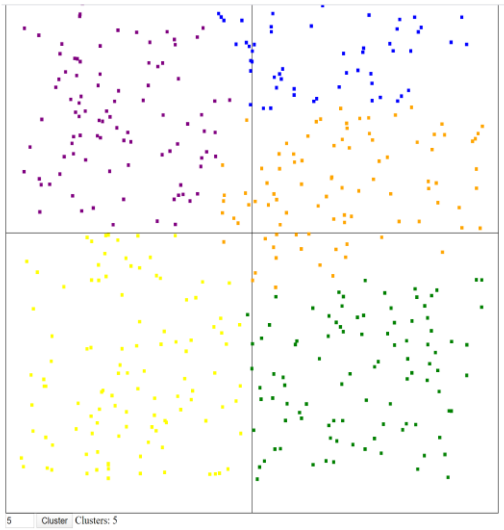


Box plots

Z-score

Scatter plot





Output

1 Threshold of Air quality value that indicates an outlier

2 A lower threshold and number of times that exceeding this threshold indicate an outlier

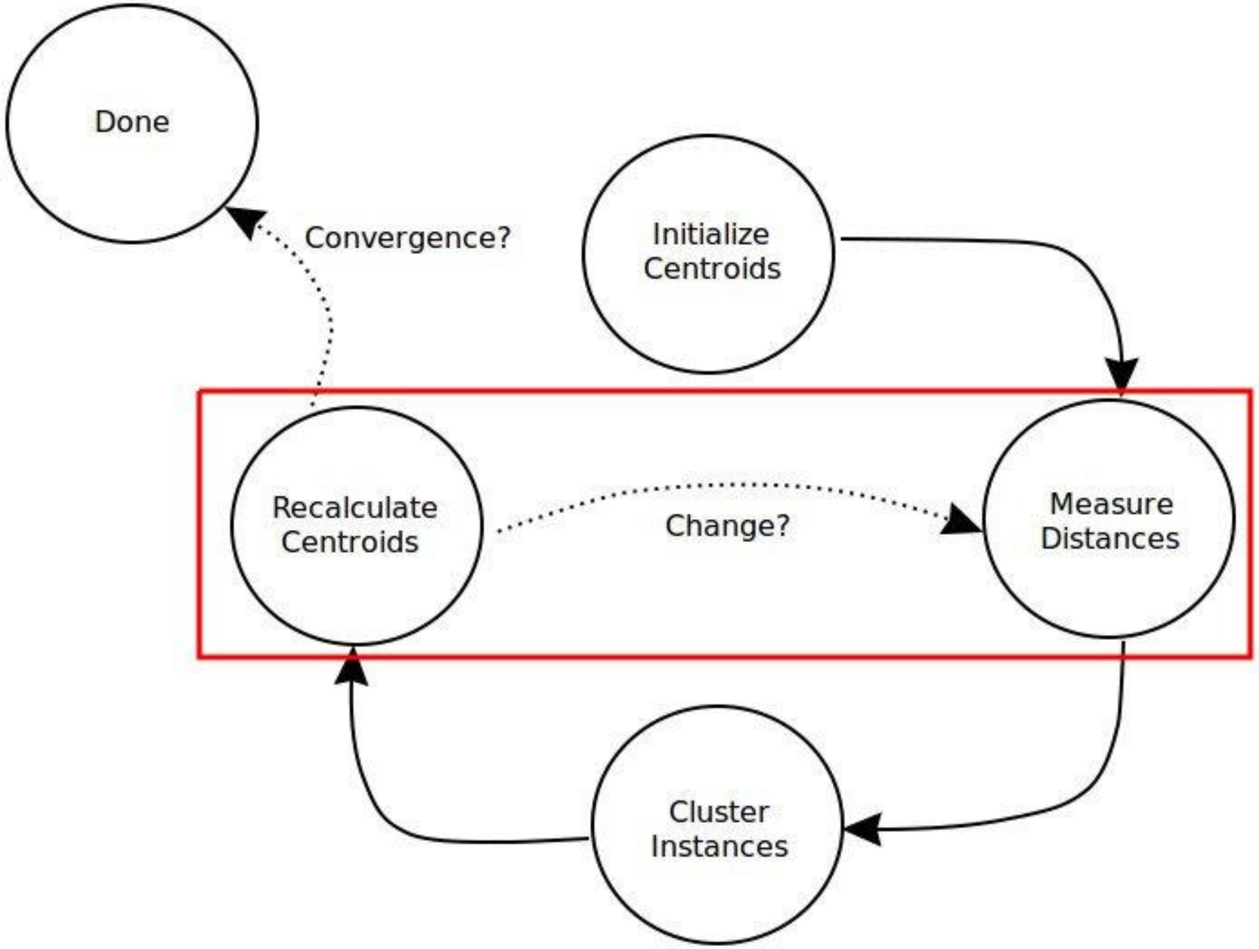
Import Air Quality dataset from OpenAQ for 3 – 6 years + more years

Python: pandas, numpy, matplotlib, sci-kit learn

1 POTENTIAL OTHER DATASETS

- Wind Direction
- Wind Speed
- Humidity
- Weather
- Rainfall
- Land Use
- Cloud cover
- Temperature

2 EVOLVING OF CLUSTER ANALYSIS



Source: https://www.kdnuggets.com/2017/08/comparing-distance-measurements-python-scipy.html

IDEAS FOR ANALYSIS

- Isochrone

The spatial distribution that affects Air Quality Sensors

How

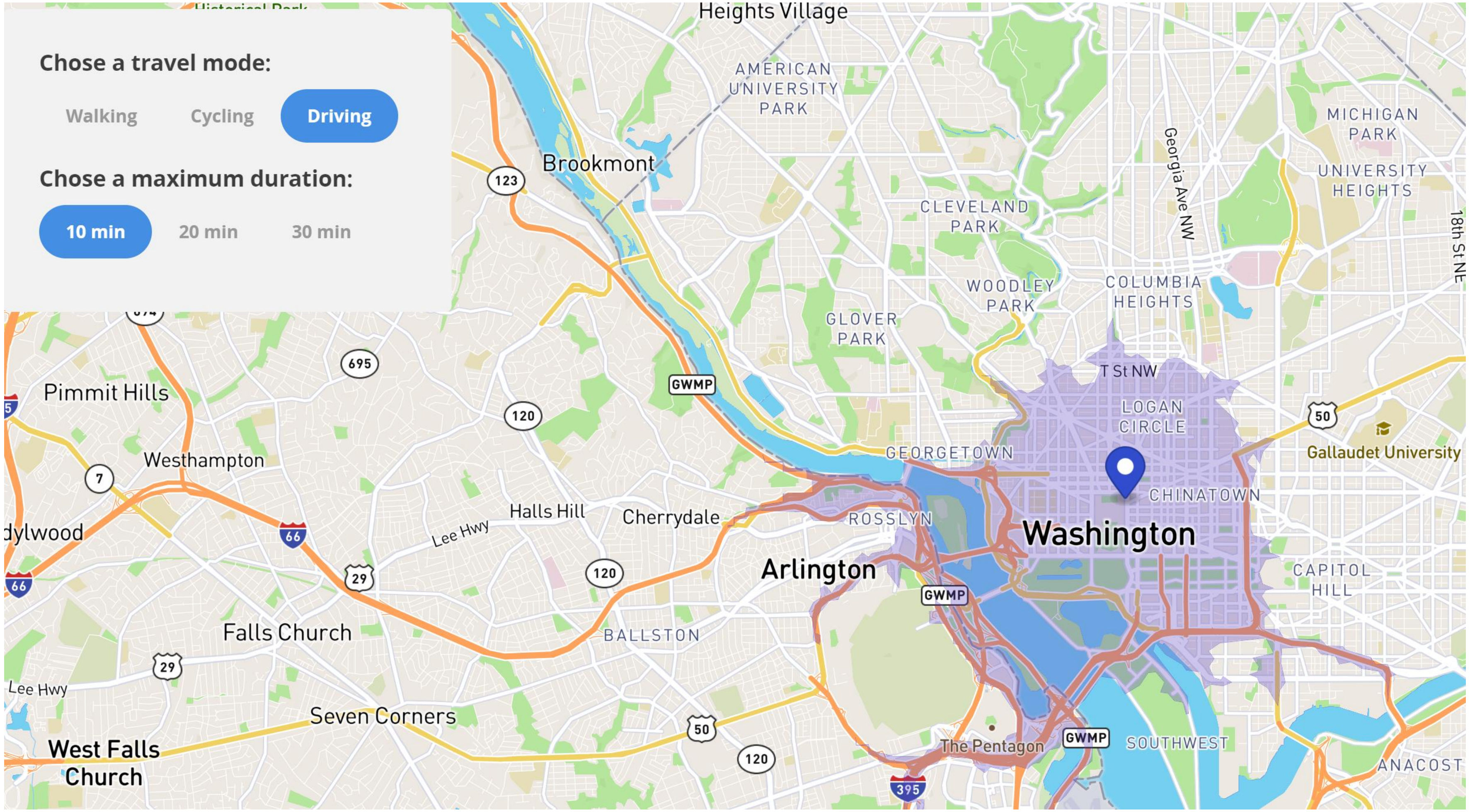
1 Use Wind direction and Speed

2 Find any obvious Emissions sources in distribution

3 Find an obstruction in distribution

(Find Wind direction where begins)

(Estimate potential distribution)



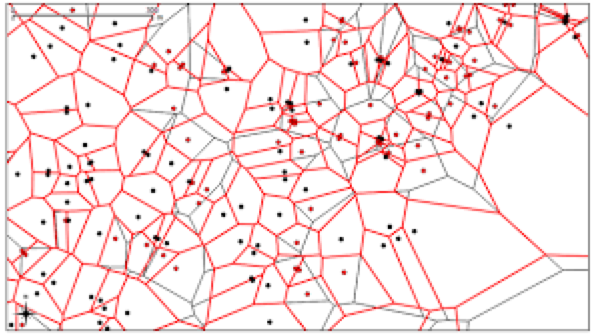
Chose a travel mode:

Walking Cycling [✓ Driving]

Chose a maximum duration:

[✓ 10 min] 20 min 30 min

SPATIAL DATASETS



- Using Voronoi Spatial Analysis for distance from other Sensors
- **Advantage**
- Distance to other OpenAQ sensors and reliability of comparing to other sensors
- Coverage of that is near to that OpenAQ sensor

Deliverables to Github

1 Presentation
2 Access to EWC and loaded anaconda
3 Jupyter Notebook of 8 Clustering Analysis
4 Revised Gantt Chart
5 PCA implementation

POTENTIALLY MORE RELIABLE STATION

- US EMBASSY Sensors (Potentially more reliable)
- From Spatial Analysis (Stations with other stations close by are more reliable)
- Stations with most measurements (More measurements means identifying frequent errors in sensors)
- Cluster similarity metrics (How formed are the clusters and whether there are many outliers)
- PCA Completed and available on the EWC (What are the most significant factors)