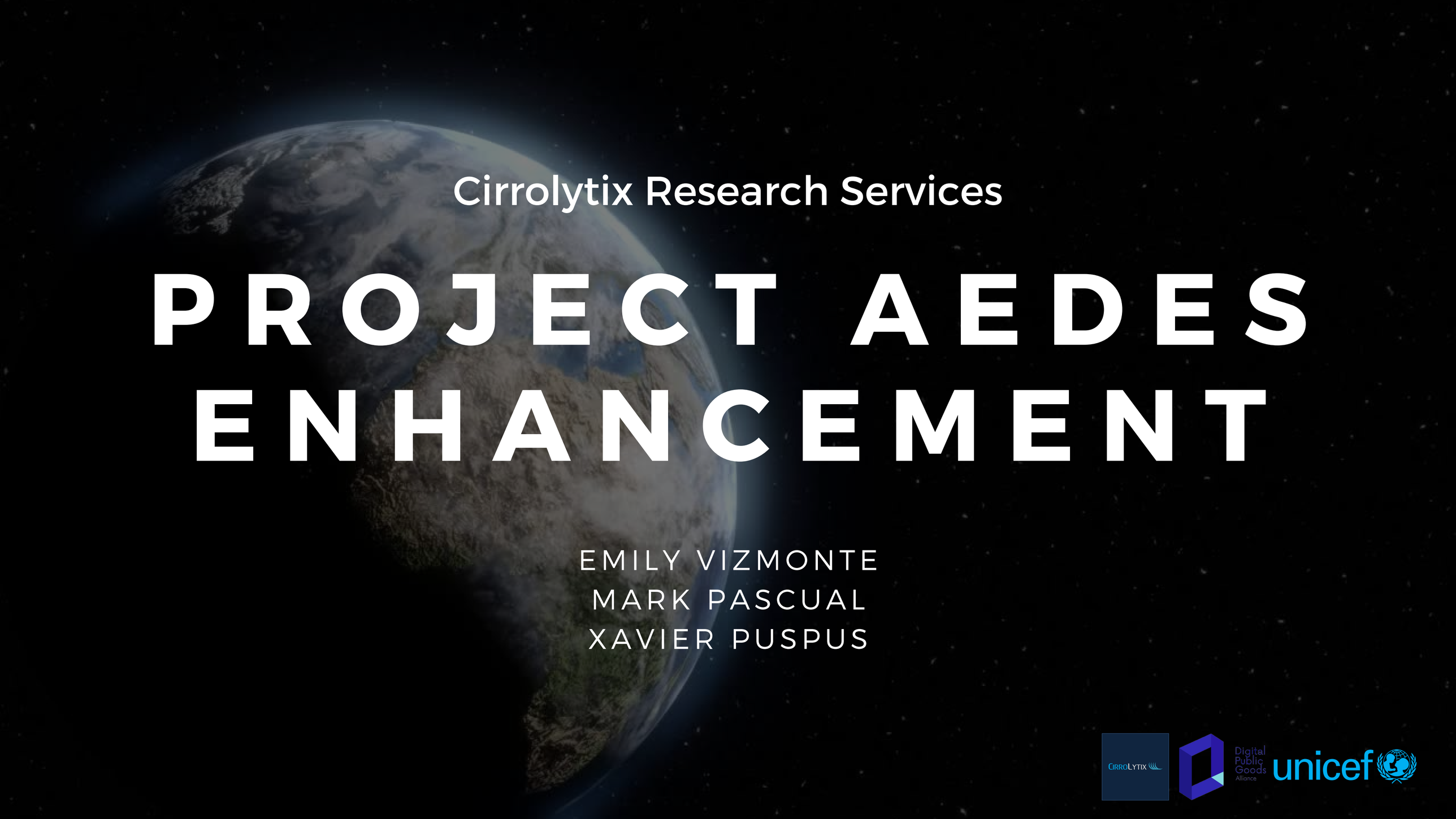


Cirrolytix Research Services

PROJECT AEDES ENHANCEMENT

EMILY VIZMONTE
MARK PASCUAL
XAVIER PUSPUS

CIRROLYTIX

Digital Public Goods Alliance

unicef

Solutions and Enhacements

PROJECT AEDES OPEN PLATFORM

FORECASTING

Predict future number of cases/deaths of vector-borne diseases

HOTSPOT DETECTION

Identify locations of possible hotspots for outbreaks

INFORM RISK MAPPING

Map out risk framework using environmental data

OPEN API

Publicly open pre-processed satellite, weather, socioeconomic and health datasets

PYTHON PACKAGE

Open-source tools used for data collection, feature engineering and automated machine learning

Status Report

Improvements	Status
Automation of data gathering from various sources.	A working Python package is now available. It allows automated data collection covering the entire data stack. It also provides the capability of visualizing all the points of interest with their proper labels using one line of code.
Addition of new weather, satellite, geospatial and socioeconomic data to enrich dataset	Data stack has been enriched with new datasets. Refer to Data Stack section.
Open API	Data Management System (CKAN) has been created and being tested to include datasets generated from the Python package.
Enhancing the predictive modeling by adding additional ML algorithms to improve model fitting performance applicable to Dengue Forecasting and Hotspot Detection	Ongoing Machine Learning (ML) Model Training. Refer to Models and Frameworks section for more details.
Incorporating the INFORM Epidemic Risk Framework with data gathered by AEDES teams to generate location-based risk maps, and advise policy interventions to mitigate the impacts of dengue	Ongoing INFORM Risk model testing utilizing current datasets.
Improvement of User Interface to make it feel more like a consumer utility.	Wireframe currently being developed.

RISK-BASED ASSESSMENT FRAMEWORK

HAZARDS

Monitor progress of epidemic, Generate alerts

VULNERABILITIES

Prioritize areas with vulnerable groups, suggest demographic and geographic determinants of risk

LACK OF COPING CAPACITY

Prioritize areas for emergency aid, recommend infrastructure investment

IDEAL RISK-BASED ASSESSMENT FRAMEWORK

HAZARDS

Dengue Case incidence
Flood Occurrence
Temperature
Precipitation
COVID-19 Incidence
Access to water
Access to sanitation

VULNERABILITIES

Population ages 0-20
Poverty Index
Population affected by natural disasters
Population previously infected by dengue
Mortality
Land-use types
Social listening
Primary and secondary schools
PhilHealth coverage
Human mobility

LACK OF COPING CAPACITY

Presence of health centers
Presence of hospitals
Number of health workers
Health expenditure
Vaccination coverage

DPG RISK-BASED ASSESSMENT FRAMEWORK

HAZARDS

Dengue Case
Temperature
Precipitation Rate
Relative Humidity
Surface Temperature
Soil Moisture Index
Water Index
Distance and count of water sources
Distance and count of sanitation and waste facilities

VULNERABILITIES

Vegetation Index
Built-up Index
Aerosol Index
Distance and count of kindergartens
Distance and count of schools
Distance and count of colleges
Distance and count of universities

LACK OF COPING CAPACITY

Distance and count of clinics
Distance and count of hospitals
Distance and count of doctors

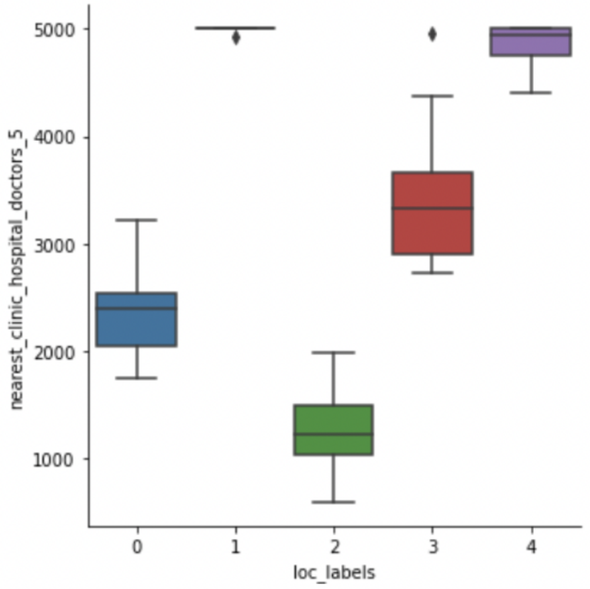
INFORM risk modeling using unsupervised learning

Clusters are relabelled into risk levels by **analyzing the distributions of each feature vs the cluster labels**

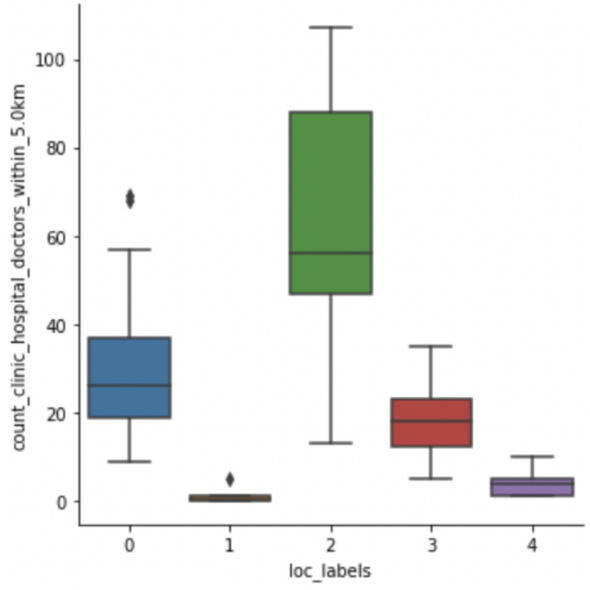
```
 # Perform clustering (this model can be saved and re-loaded later using joblib)
loc_model = perform_clustering(df,
                               features=loc_features,
                               n_clusters=5)
# Create the labels
loc_labels = pd.Series(loc_model.labels_)

# Create INFORM risk dataframe for Lack of Coping Capacity
loc_full_df = df[loc_features].drop_duplicates()
loc_full_df['loc_labels'] = loc_labels
```

We then perform analysis on categorical comparison of feature distribution as demonstrated by the sample images below:

Nearest Clinics/Hospitals

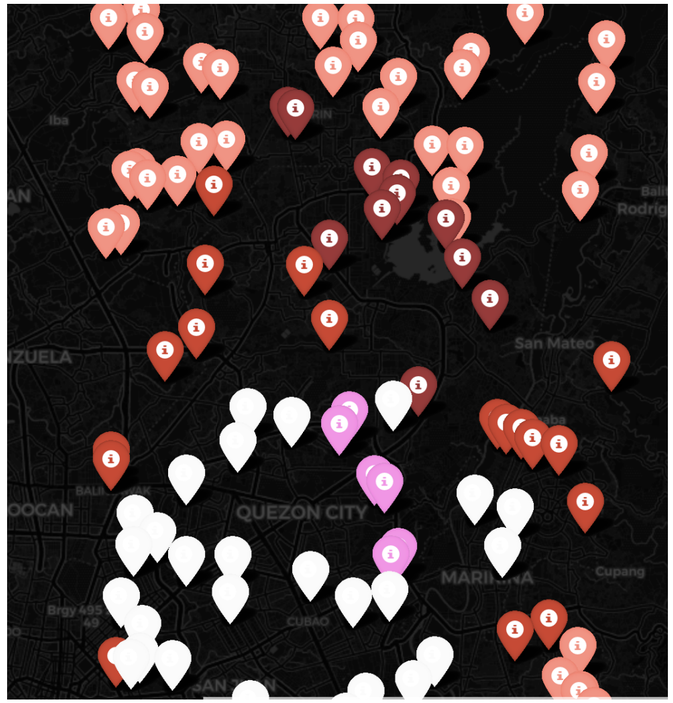


Count of Clinics/Hospitals



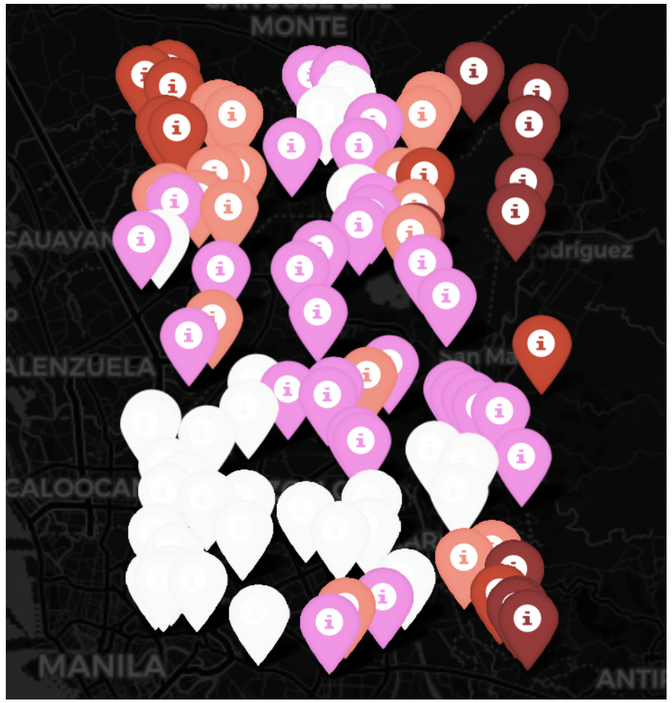
INFORM Risk Maps

Clustering is automated but risk level identification is identified using **analysis on categorical comparison of feature distributions**



Hazards

Vulnerabilities

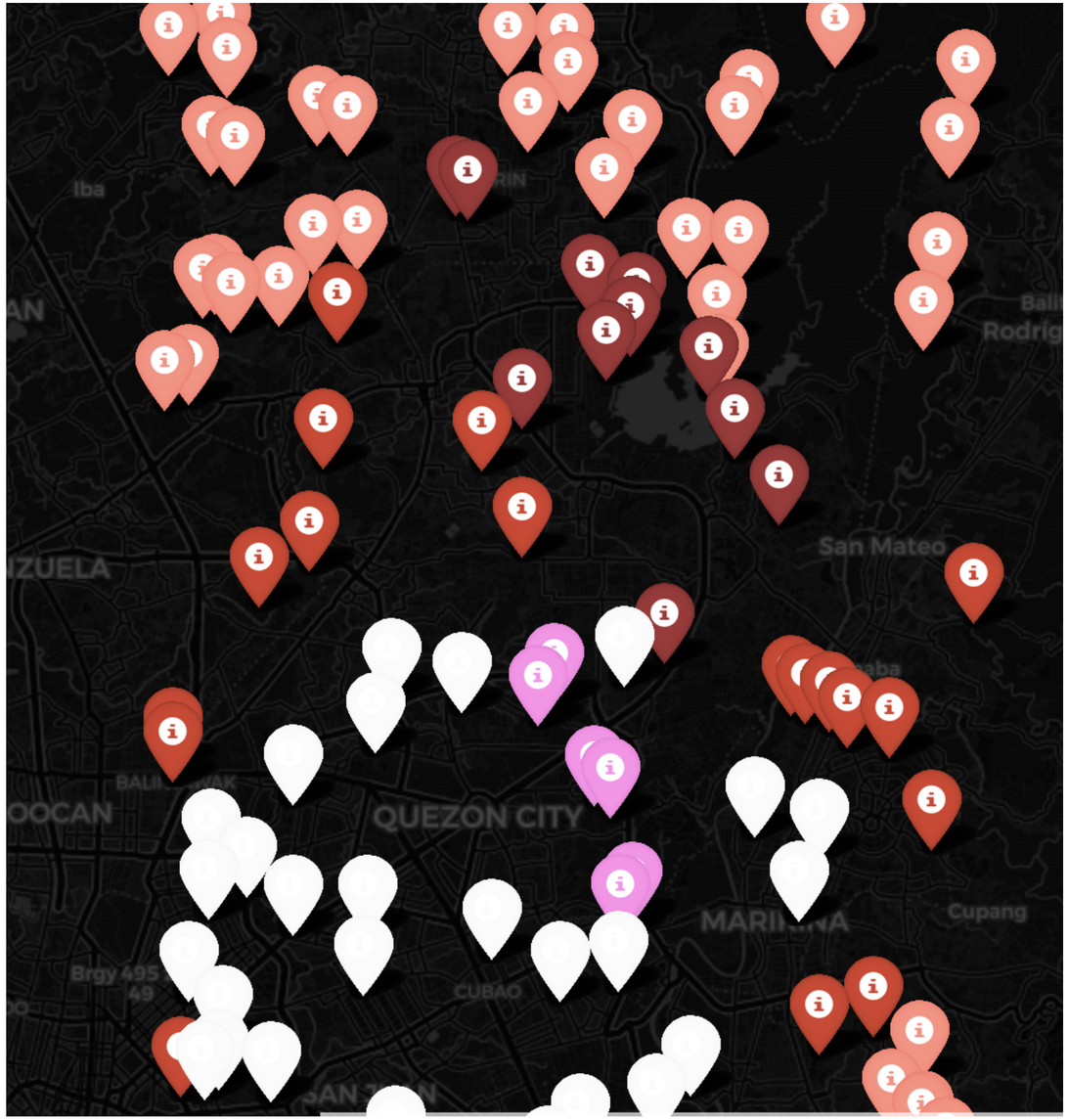


Lack of Coping Capacity

Hazards

Using geospatial clustering, the model automatically identifies locations with **high count of predicted dengue outbreaks, high water index, high count and near to toilets and water points**

Darker is riskier



Vulnerabilities

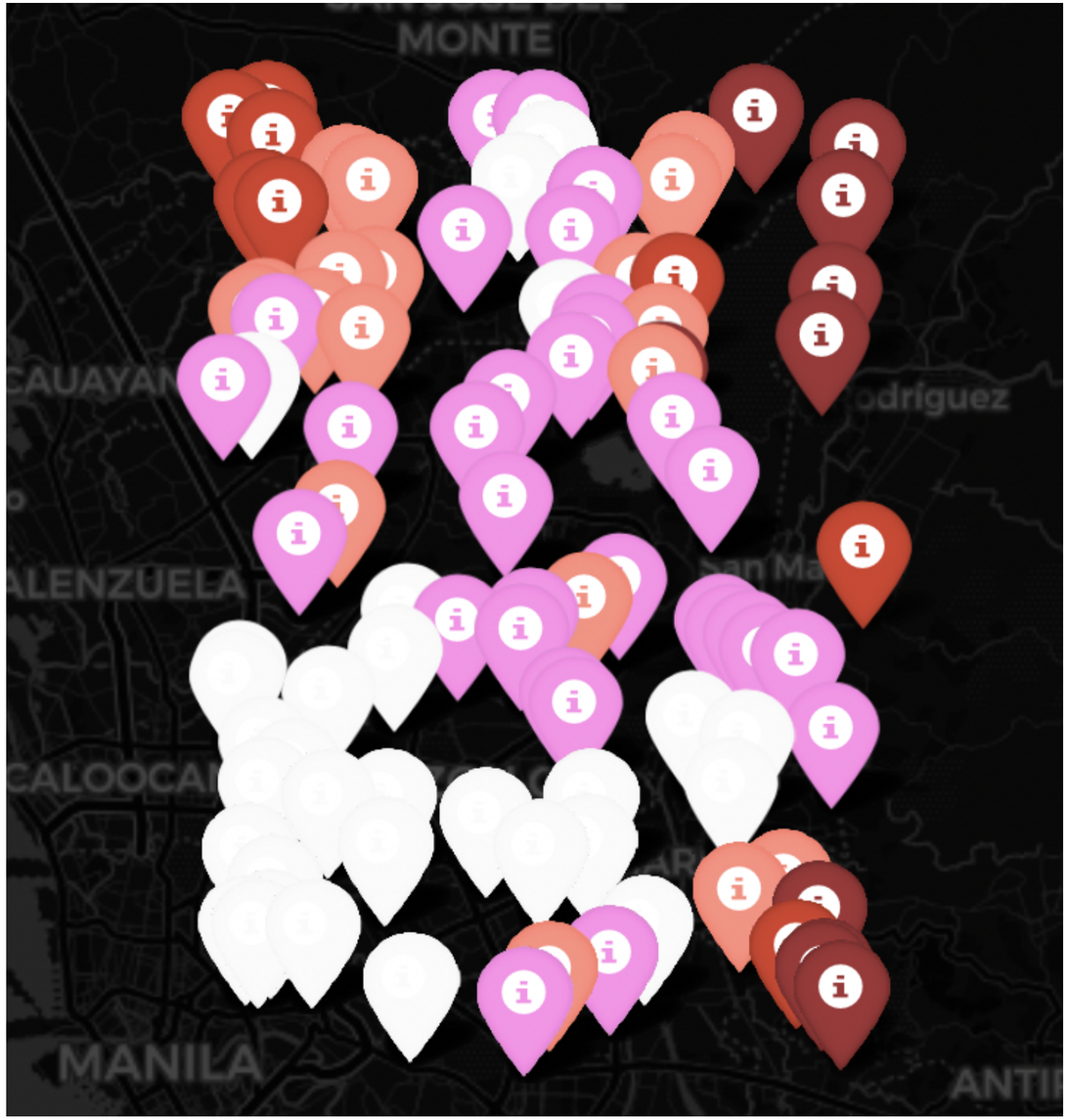
Using geospatial clustering, the model automatically identifies locations with **high urbanicity, high vegetation and high count and close to schools (kindergartens to universities)**

Darker is riskier

Lack of Coping Capacity

Using geospatial clustering, the model automatically identifies locations with **low count of or far from clinics and hospitals**

Darker is riskier



FRONT-END|BASICS

FRONT-END|FRAMEWORKS



FRONT-END|TOOLS

BACK-END


DATABASE | RDBMS

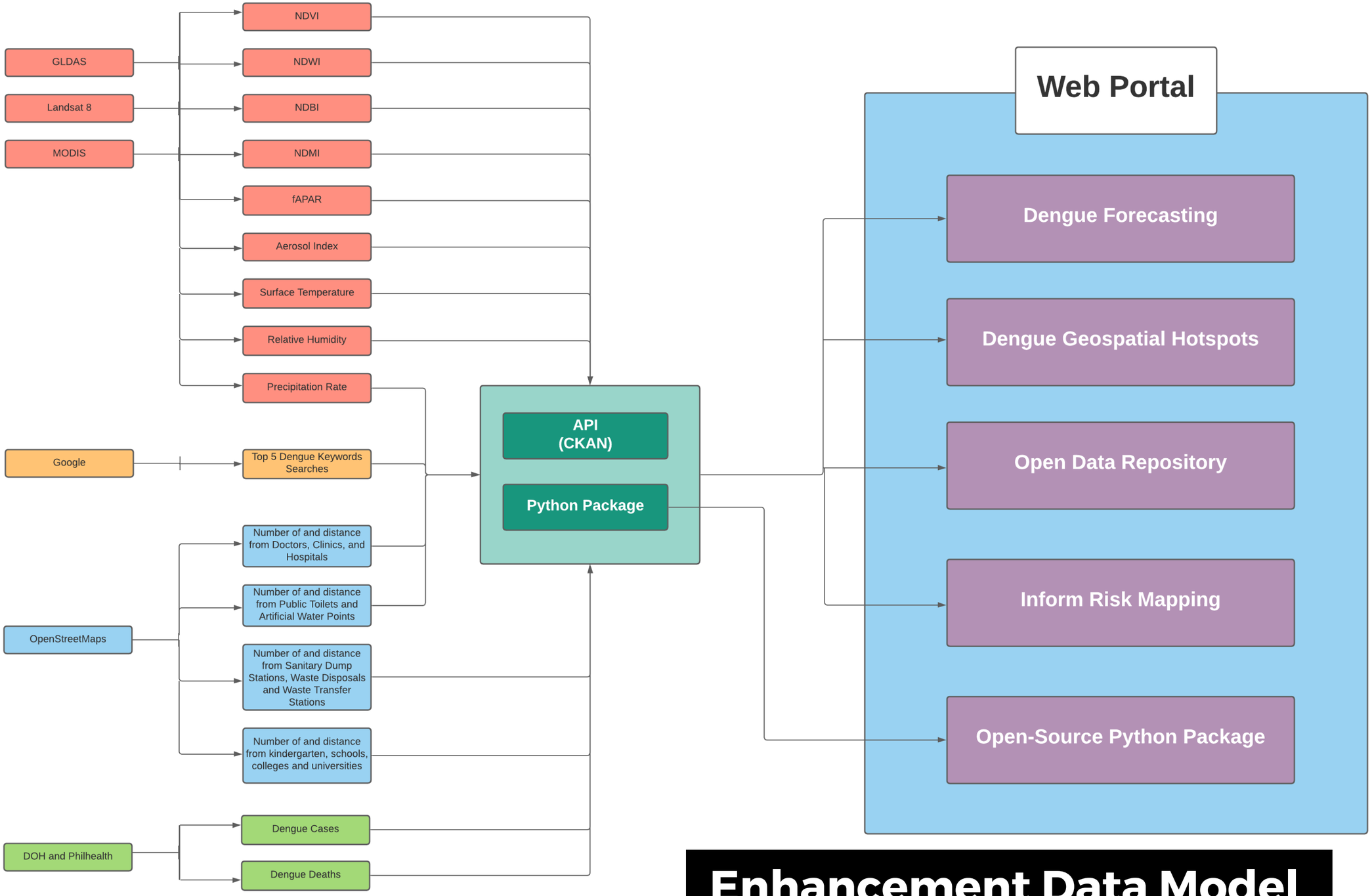
DEV-OPS|INFRASTRUCTURES

DEV-OPS|DEVELOP

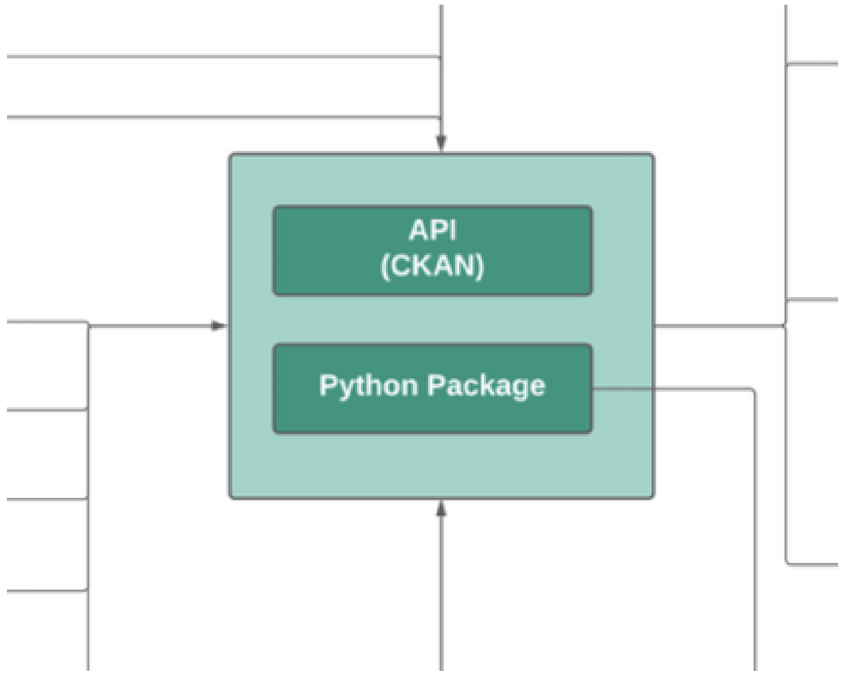
DATA SERVICES


DATA MANAGEMENT SYSTEM

ENHANCED TECH STACK



Enhancement Data Model



Data Management System

Python Package

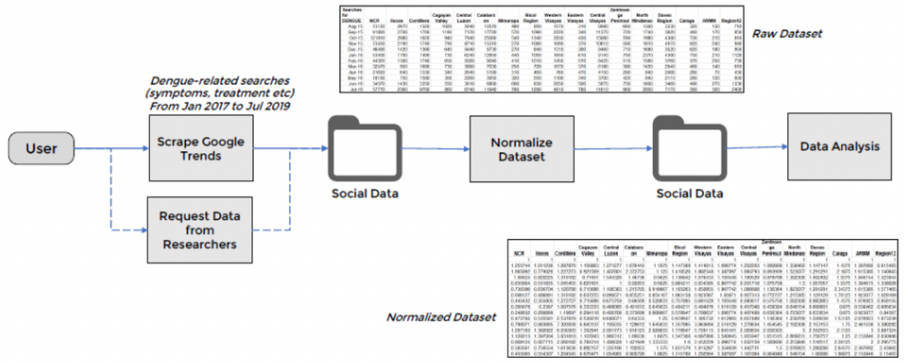
Data Collection Process Prior to Enhancement

DATA COLLECTION WAS MANUAL AND ALL OVER THE PLACE

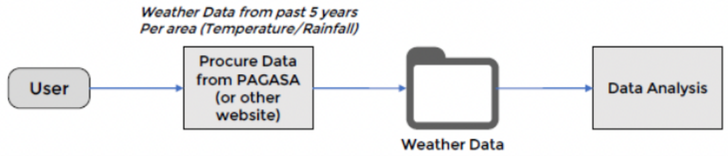
Previous Process

The following diagrams show the existing data gathering processes for each data category.

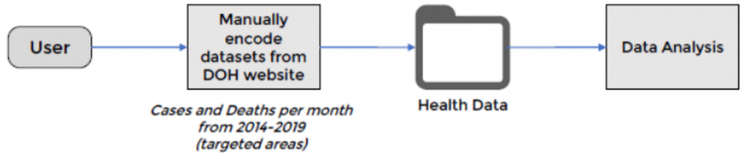
Social Listening Data



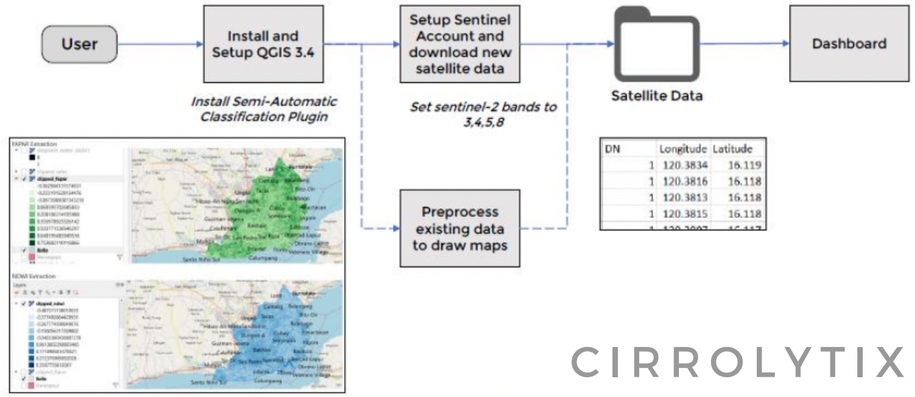
Weather Data



Health Data



Geospatial Data



Material Dashboard 2

Dashboard

Dashboard

Search here

1

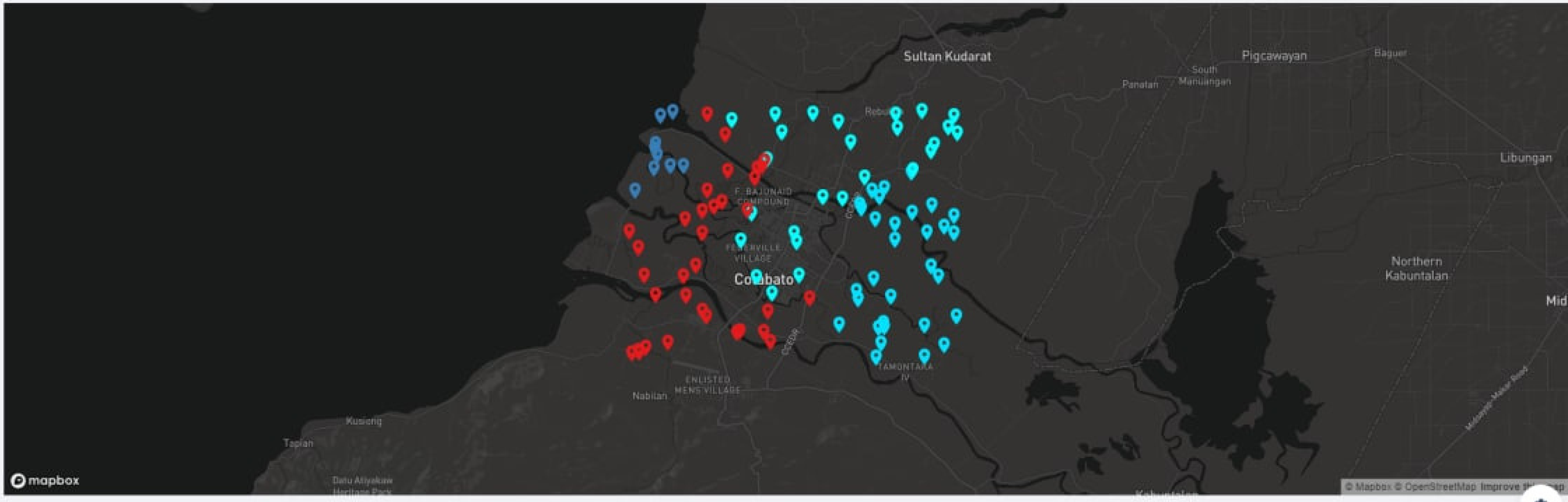
Open Data through API and Dashboard Design

https://aedes-datacatalogue-beta.xyz/

Points with Remote Sensing

ID	LONGITUDE	LATITUDE	NDVI	NDBI
1	124.23271102909972	7.2322490886052195	0.29426543772184005	-0.18135909206120807
2	124.188791024222	7.1651933675878565	0.33712953424396913	-0.17811715393663746
3	124.20985630039459	7.196758407621468	0.2525439835434753	-0.16634924621562536
4	124.19676303625263	7.2513988770380315	0.049508159867661314	-0.08317019477819493
5	124.29736612018179	7.218621620798705	0.24762520238554964	-0.1811293855035335
6	124.30661530426687	7.178069029871818	0.362507864123868	-0.2336097860415473
7	124.26744081993178	7.2416938171902725	0.3057146163256264	-0.1561891062137156
8	124.29377210327505	7.253233563803932	0.3395442103023234	-0.18009947910026575
9	124.2778049714143	7.173807474207336	0.3199756208166676	-0.2031917028784548
10	124.26291924257728	7.249400110845375	0.3182352310440857	-0.16604603124972356

1



README.md

AEDES
Python Package

PYPI: https://pypi.org/project/aedes/
repository: https://github.com/xmpuspus/aedes

Visualize on a Map

This packages also provides the capability of visualizing all the points on the map with only one line of code.

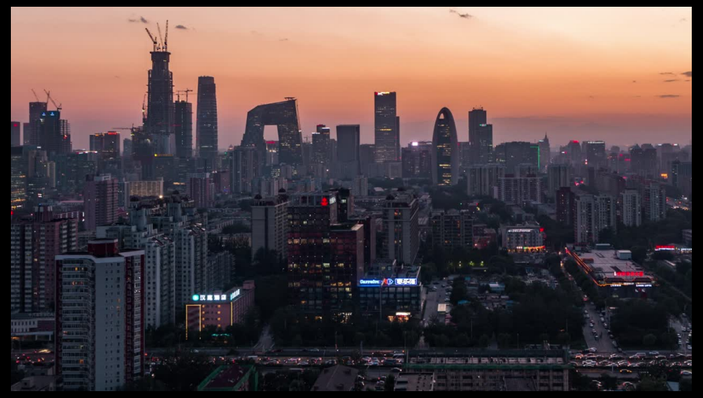
```
vizo = visualize_on_map(rev_geocode_qc_df)
vizo
```

Dengue Detection Features

COLLECTING SATELLITE DATA AT POINTS OF INTEREST

Vegetation and Water

NDWI
NDVI
FAPAR
NDMI



Urbanicity

NDBI
Aerosol Index

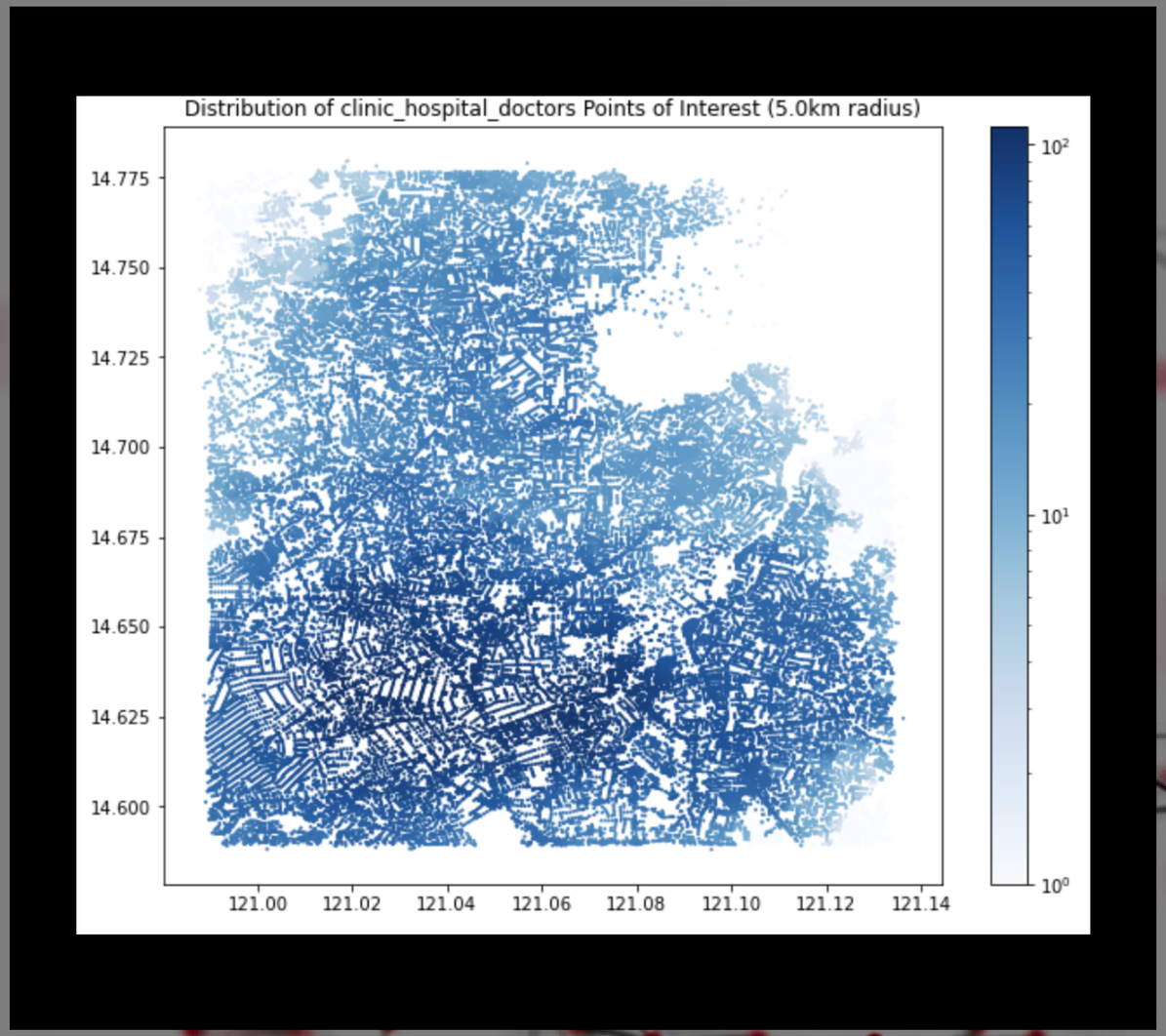
Weather

Surface Temperature
Precipitation Rate
Relative Humidity

Heathcare Capacity

DISTRIBUTION PER RADIAL BLOCK

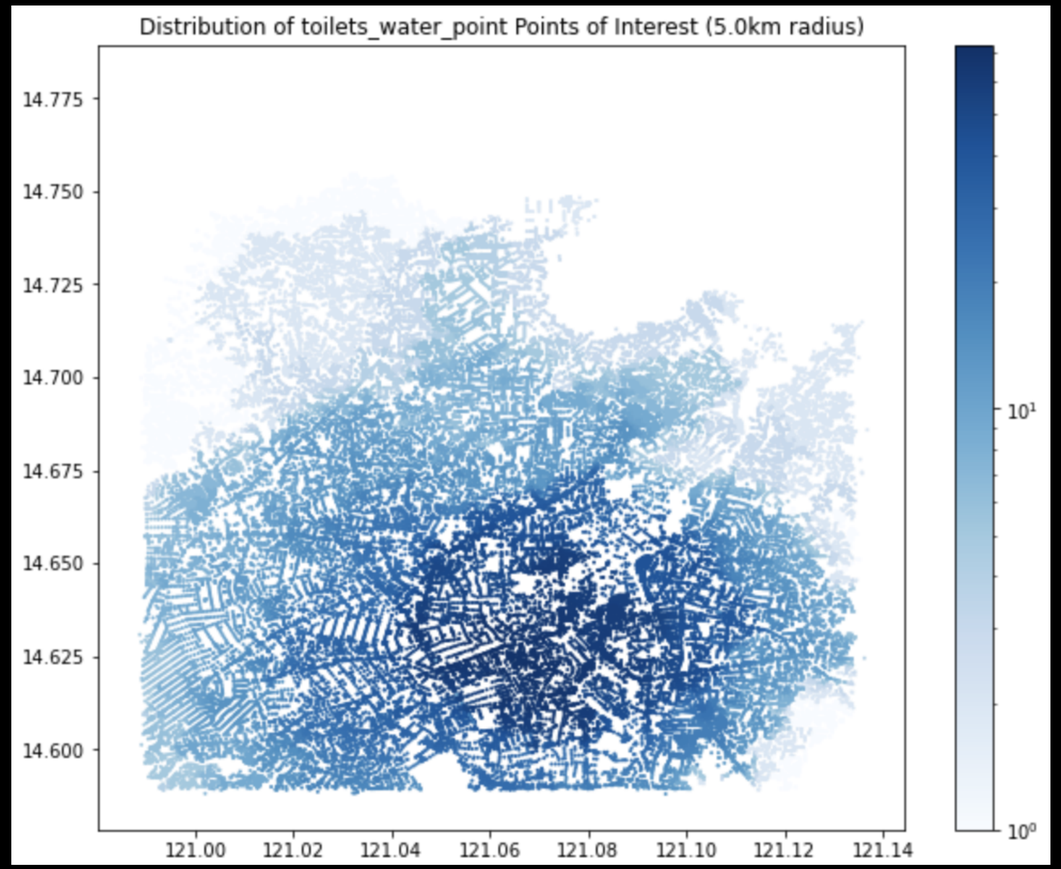
Heatmap of **accessibility to nearest clinic, and hospital**



Access to Water Sources

DISTRIBUTION PER RADIAL BLOCK

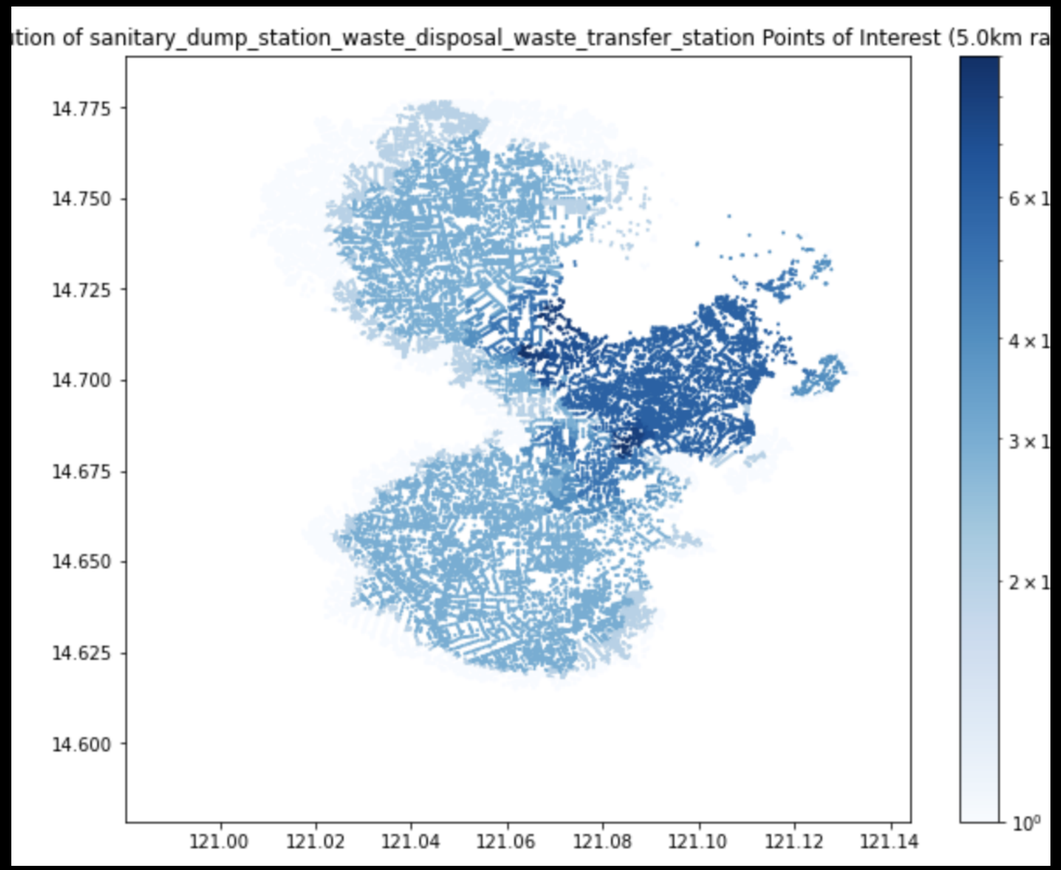
Heatmap of **accessibility to nearest toilet facility and different types of water points as defined by OSM**



Access to Sanitation

DISTRIBUTION PER RADIAL BLOCK

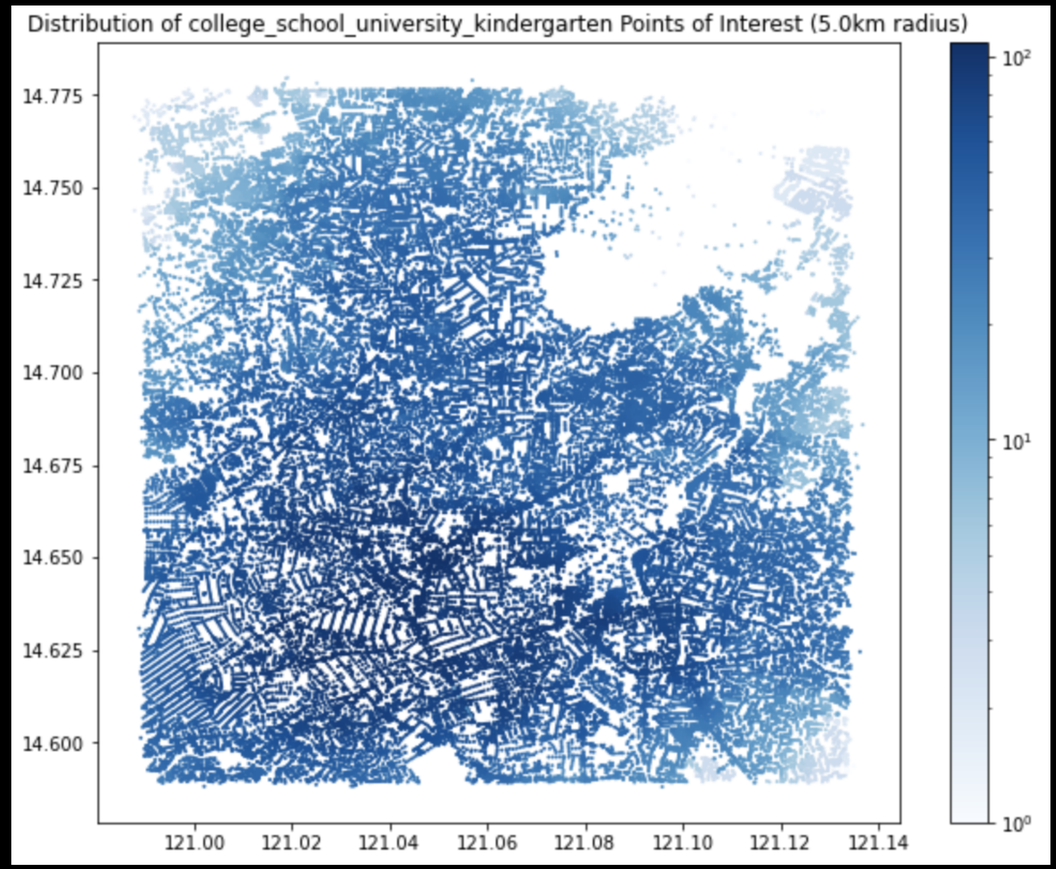
Heatmap of **accessibility to nearest sanitary dump station, waste disposal, and waste transfer stations**



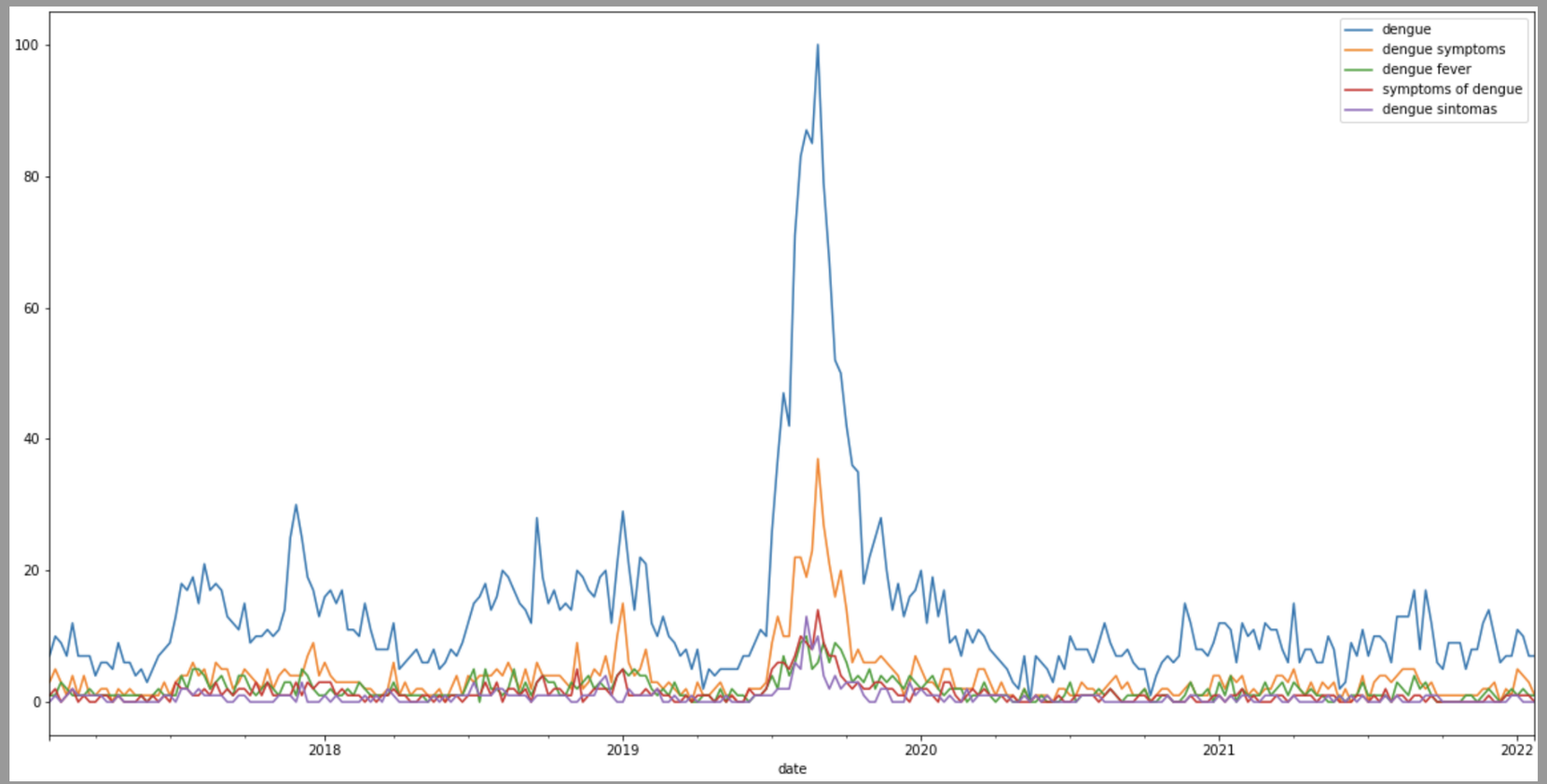
Schools

DISTRIBUTION PER RADIAL BLOCK

Heatmap of **accessibility to nearest kindergarten, schools, colleges, and universities**

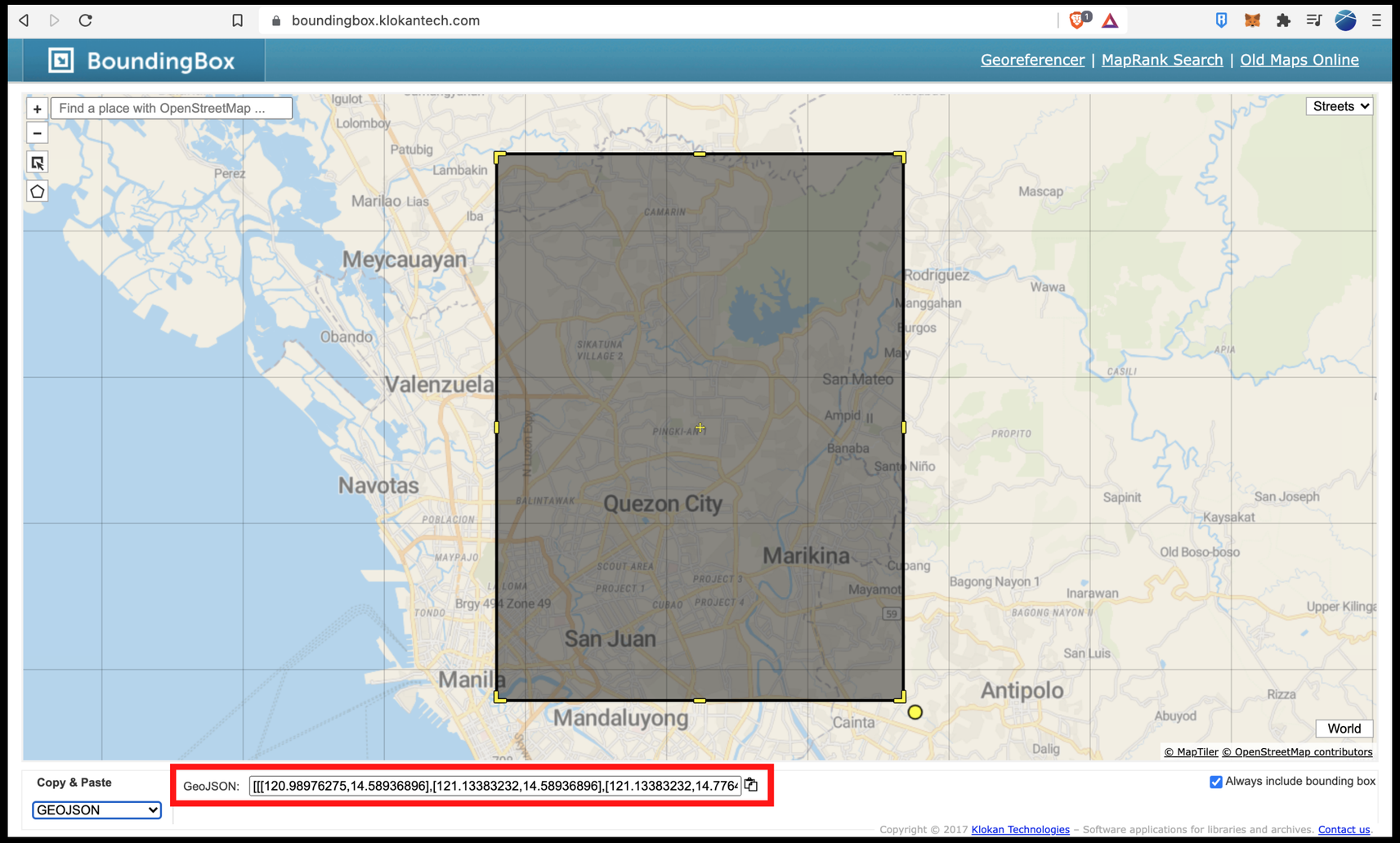


Google Search Trends



Code	Subdivision name (en)
PH-14	Autonomous Region in Muslim Mindanao[b]
PH-05	Bicol
PH-02	Cagayan Valley
PH-40	Calabarzon
PH-13	Caraga
PH-03	Central Luzon
PH-07	Central Visayas
PH-15	Cordillera Administrative Region
PH-11	Davao
PH-08	Eastern Visayas
PH-01	Ilocos
PH-41	Mimaropa
PH-00	National Capital Region
PH-10	Northern Mindanao
PH-12	Soccsksargen
PH-06	Western Visayas
PH-09	Zamboanga Peninsula

Sample Web App Data Input



In order to collect the data and features necessary for modelling, the only input is a **bounding box geojson or a polygon**

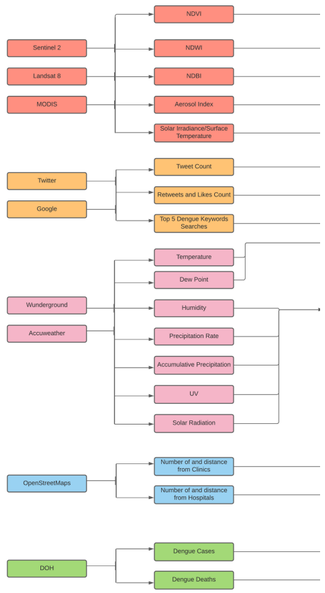
Extracted Data Features

```
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   geometry                                                              50 non-null     geometry
 1   buffered_geometry                                                     50 non-null     object
 2   longitude                                                             50 non-null     float64
 3   latitude                                                              50 non-null     float64
 4   ndvi                                                                  50 non-null     float64
 5   fapar                                                                 50 non-null     float64
 6   ndbi                                                                  50 non-null     float64
 7   ndwi                                                                  50 non-null     float64
 8   ndmi                                                                  50 non-null     float64
 9   aerosol                                                               50 non-null     float64
 10  surface_temperature                                                   50 non-null     float64
 11  precipitation_rate                                                    50 non-null     float64
 12  relative_humidity                                                     50 non-null     float64
 13  labels                                                                50 non-null     int32
 14  OSM_network_id                                                        50 non-null     int64
 15  nearest_clinic_hospital_doctors_1                                     50 non-null     float64
 16  nearest_clinic_hospital_doctors_2                                     50 non-null     float64
 17  nearest_clinic_hospital_doctors_3                                     50 non-null     float64
 18  nearest_clinic_hospital_doctors_4                                     50 non-null     float64
 19  nearest_clinic_hospital_doctors_5                                     50 non-null     float64
 20  count_clinic_hospital_doctors_within_5.0km                            50 non-null     float64
 21  nearest_toilets_water_point_1                                         50 non-null     float64
 22  nearest_toilets_water_point_2                                         50 non-null     float64
 23  nearest_toilets_water_point_3                                         50 non-null     float64
 24  nearest_toilets_water_point_4                                         50 non-null     float64
 25  nearest_toilets_water_point_5                                         50 non-null     float64
 26  count_toilets_water_point_within_5.0km                                50 non-null     float64
 27  nearest_sanitary_dump_station_waste_disposal_waste_transfer_station_1 50 non-null     float64
 28  nearest_sanitary_dump_station_waste_disposal_waste_transfer_station_2 50 non-null     float64
 29  nearest_sanitary_dump_station_waste_disposal_waste_transfer_station_3 50 non-null     float64
 30  nearest_sanitary_dump_station_waste_disposal_waste_transfer_station_4 50 non-null     float64
 31  nearest_sanitary_dump_station_waste_disposal_waste_transfer_station_5 50 non-null     float64
 32  count_sanitary_dump_station_waste_disposal_waste_transfer_station_within_5.0km 50 non-null float64
 33  nearest_college_school_university_kindergarten_1                      50 non-null     float64
 34  nearest_college_school_university_kindergarten_2                      50 non-null     float64
 35  nearest_college_school_university_kindergarten_3                      50 non-null     float64
 36  nearest_college_school_university_kindergarten_4                      50 non-null     float64
 37  nearest_college_school_university_kindergarten_5                      50 non-null     float64
 38  count_college_school_university_kindergarten_within_5.0km             50 non-null     float64
dtypes: float64(35), geometry(1), int32(1), int64(1), object(1)
```

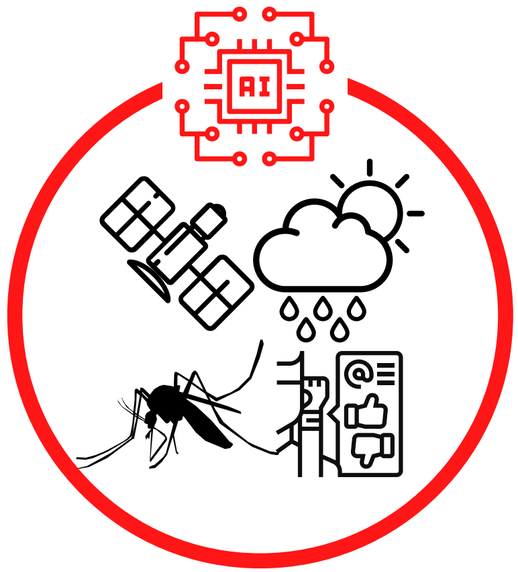
ML Model

The datasets per geolocation are feature-engineered and fed through to a machine learning model

Data

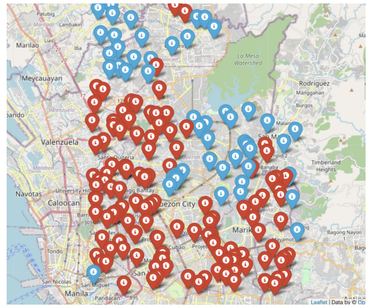


Machine Learning Model Training



Predictions

Dengue Hotspots over time

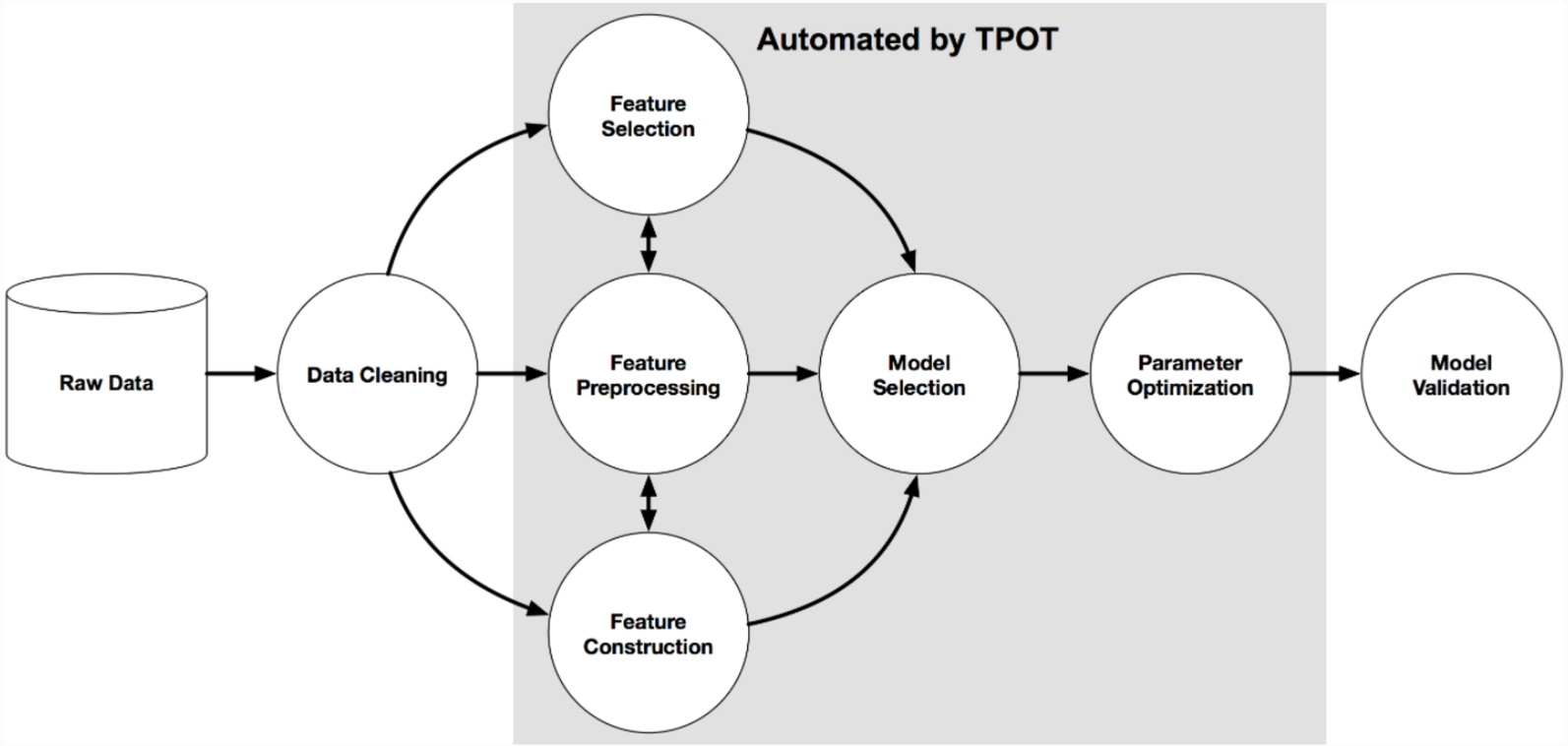


AutoML

Using Tree-based Pipeline Optimization

https://epistasislab.github.io/tpot/

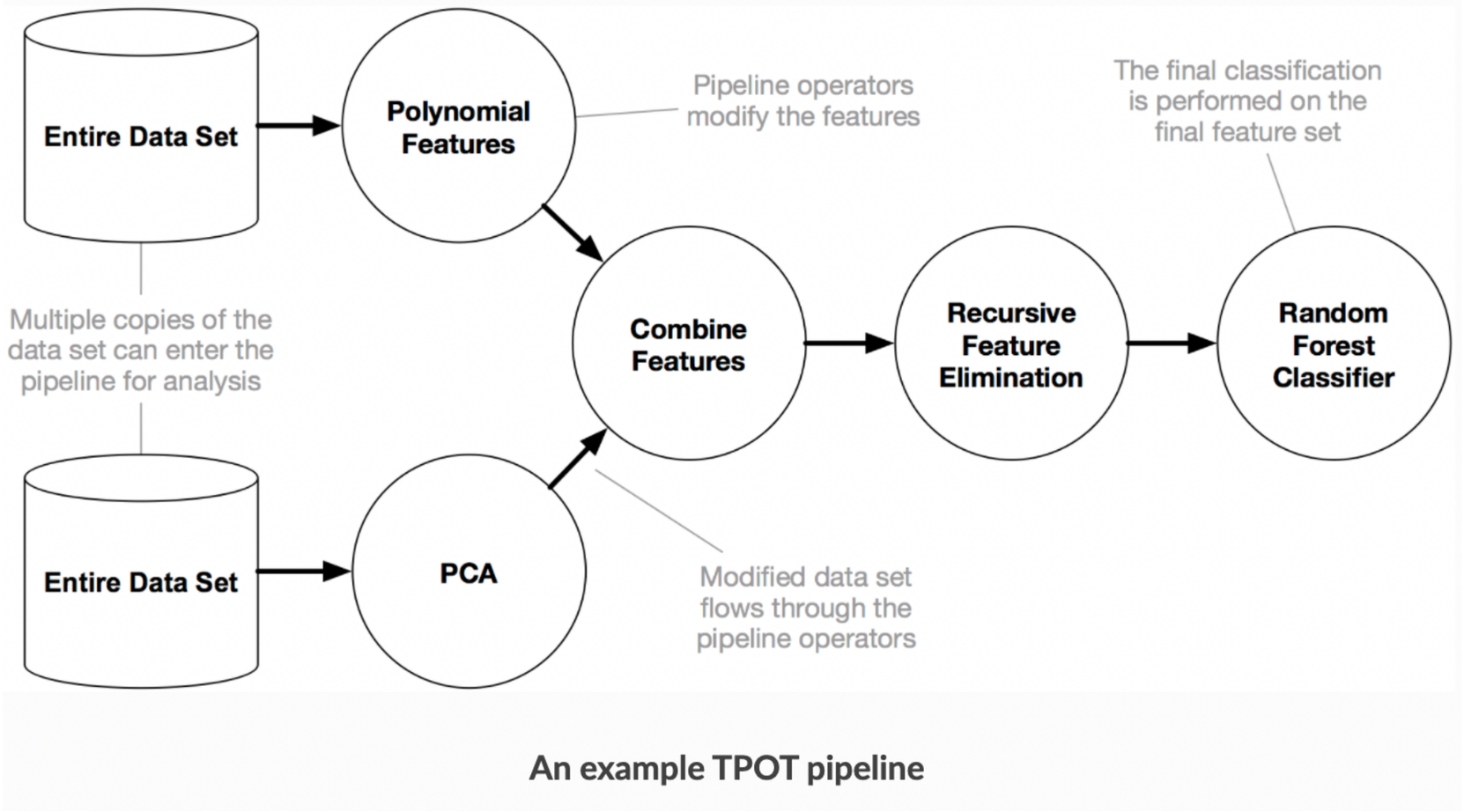
TPOT will automate the most tedious part of machine learning by intelligently exploring thousands of possible pipelines to find the best one for your data.



AutoML

Using Tree-based Pipeline Optimization

https://epistasislab.github.io/tpot/



An example TPOT pipeline

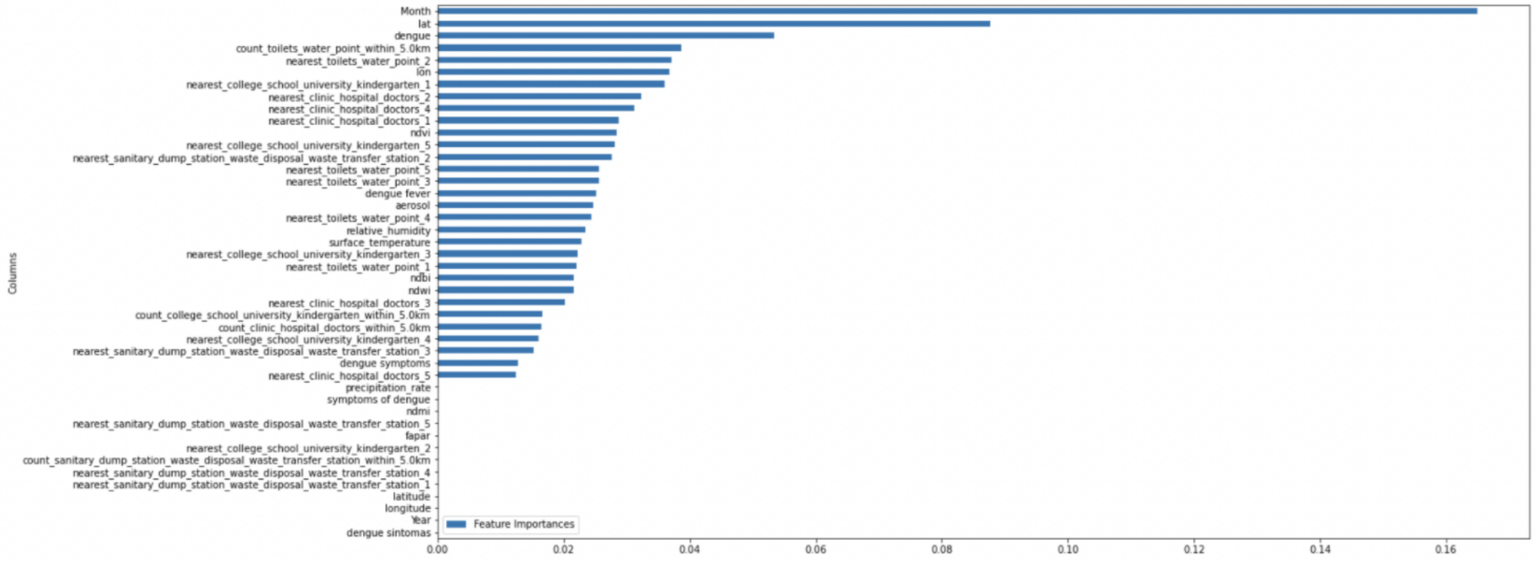
One-liner Classification Model

Aedes can also do:
perform_regression()
perform_classification()
perform_clustering()

```
model, feature_imps_df = perform_classification(X_train, y_train)
```

The output should look like this:

```
Generation 16 - Current best internal CV score: 0.889950753668092

Generation 17 - Current best internal CV score: 0.889950753668092

Generation 18 - Current best internal CV score: 0.889950753668092

Generation 19 - Current best internal CV score: 0.889950753668092

Generation 20 - Current best internal CV score: 0.889950753668092

Best pipeline: XGBClassifier(input_matrix, learning_rate=1.0, max_depth=1, min_child_weight=10, n_estimators=100, n_jobs=
1, subsample=0.6000000000000001, verbosity=0)
Best model pickle file and best model pipeline saved to the same directory as this code.
```



Automated ML Pipeline Builder

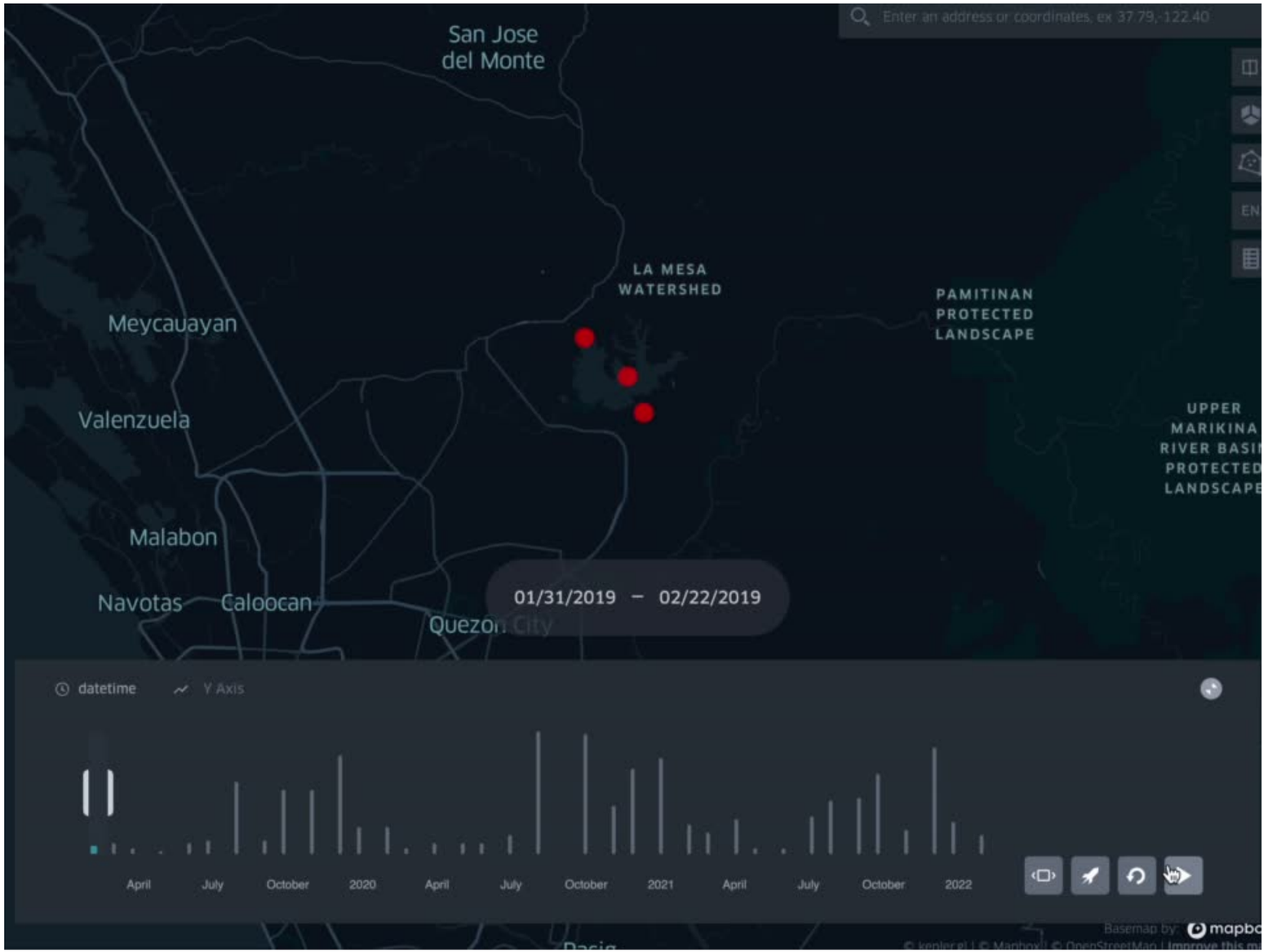
The script below is automatically written and computer-generated from one-line of code

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.impute import SimpleImputer

# NOTE: Make sure that the outcome column is labeled 'target' in the data file
tpot_data = pd.read_csv('PATH/TO/DATA/FILE', sep='COLUMN_SEPARATOR', dtype=np.float64)
features = tpot_data.drop('target', axis=1)
training_features, testing_features, training_target, testing_target = \
            train_test_split(features, tpot_data['target'], random_state=42)

imputer = SimpleImputer(strategy="median")
imputer.fit(training_features)
training_features = imputer.transform(training_features)
testing_features = imputer.transform(testing_features)

# Average CV score on the training set was: 0.889950753668092
exported_pipeline = XGBClassifier(learning_rate=1.0, max_depth=1, min_child_weight=10, n_estimators=100, n_jobs=1, subsample=0.6000000000000001, verbosity=0)
# Fix random state in exported estimator
if hasattr(exported_pipeline, 'random_state'):
    setattr(exported_pipeline, 'random_state', 42)

exported_pipeline.fit(training_features, training_target)
results = exported_pipeline.predict(testing_features)
```

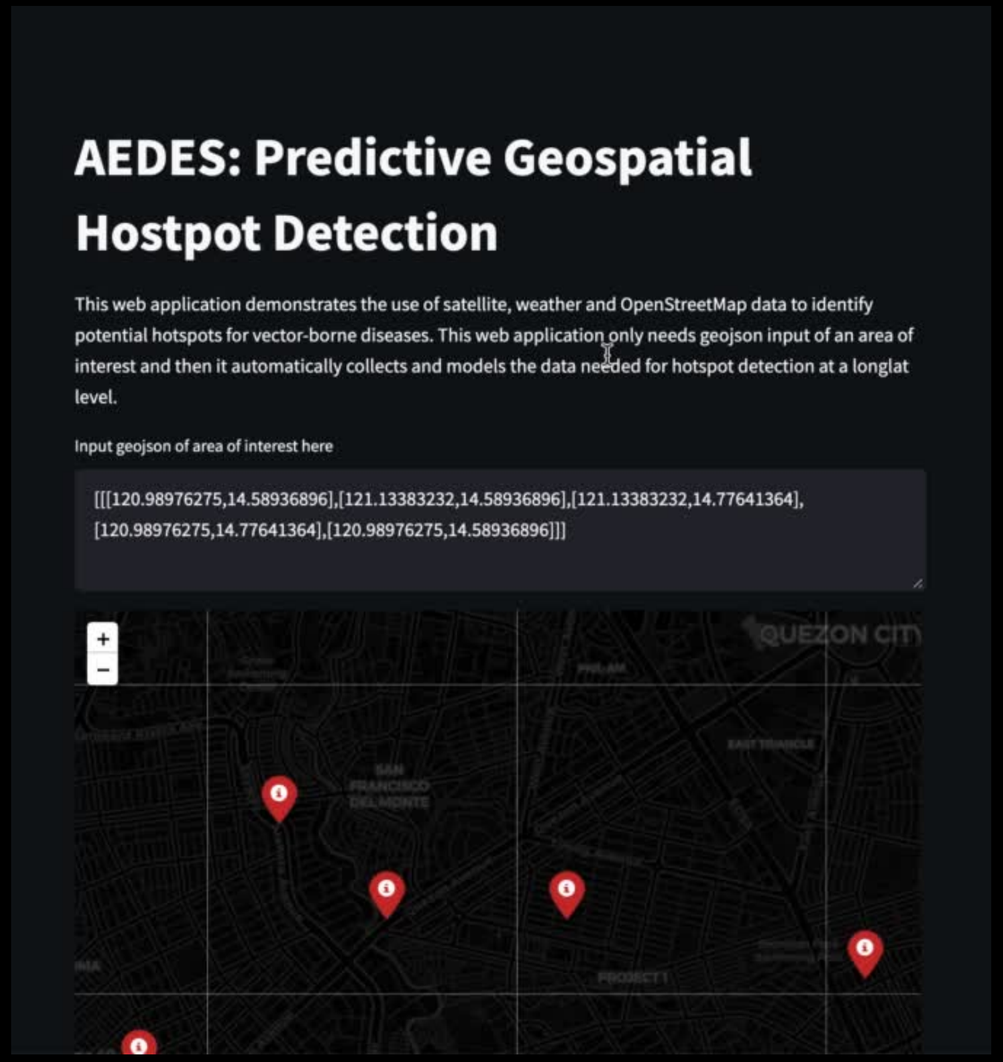
Hotspot Detections Through Time

Quezon City
2019-present

Classification Model
(over 10 dengue cases per month or not)
F1 Score:
Nowcasting model: 0.76
Month-ahead Forecasting Model: 0.74



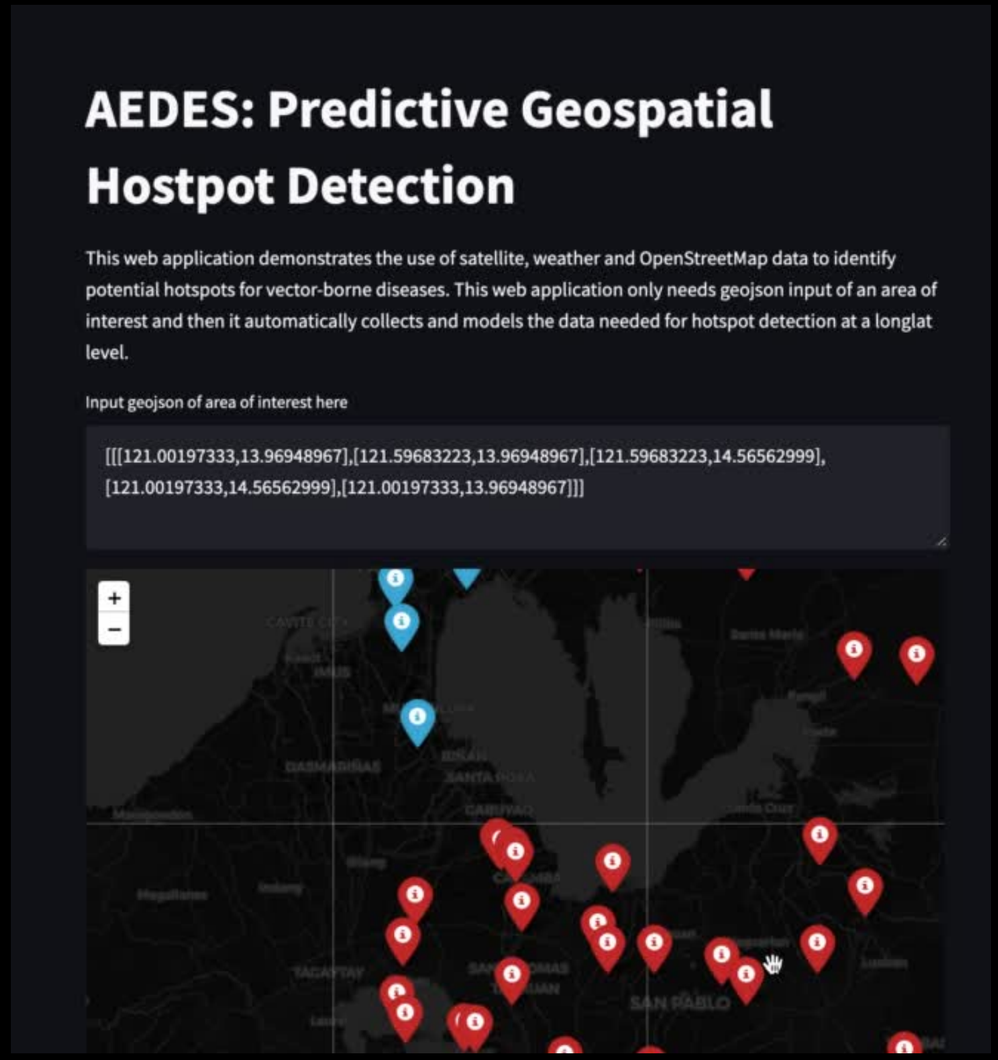
Web Application



QUEZON CITY, PHILIPPINES

From just bounding box coordinates of an area of interest, we're able to identify potential at-risk locations for dengue hotspots

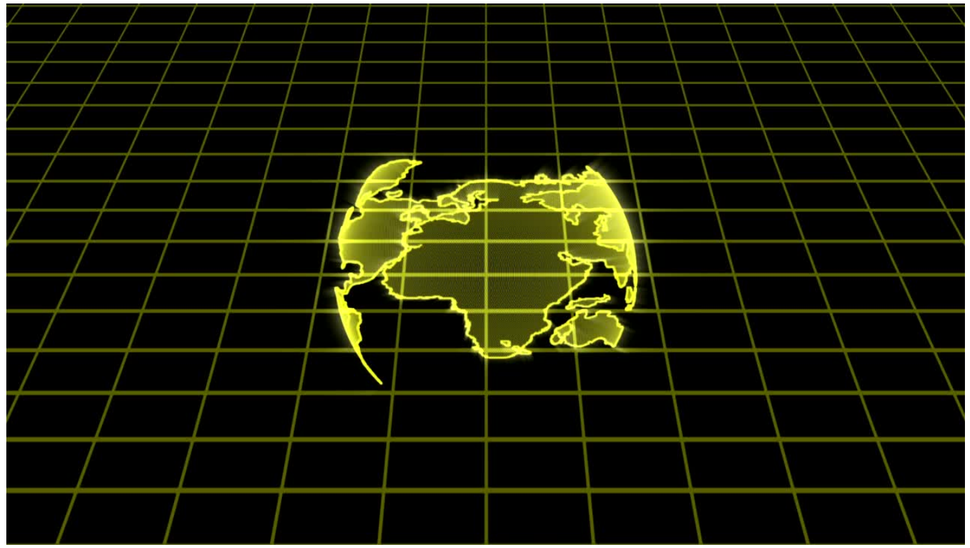
Web Application



LAGUNA, PHILIPPINES

From just bounding box coordinates of an area of interest, we're able to identify potential at-risk locations for dengue hotspots

ML Inference Input

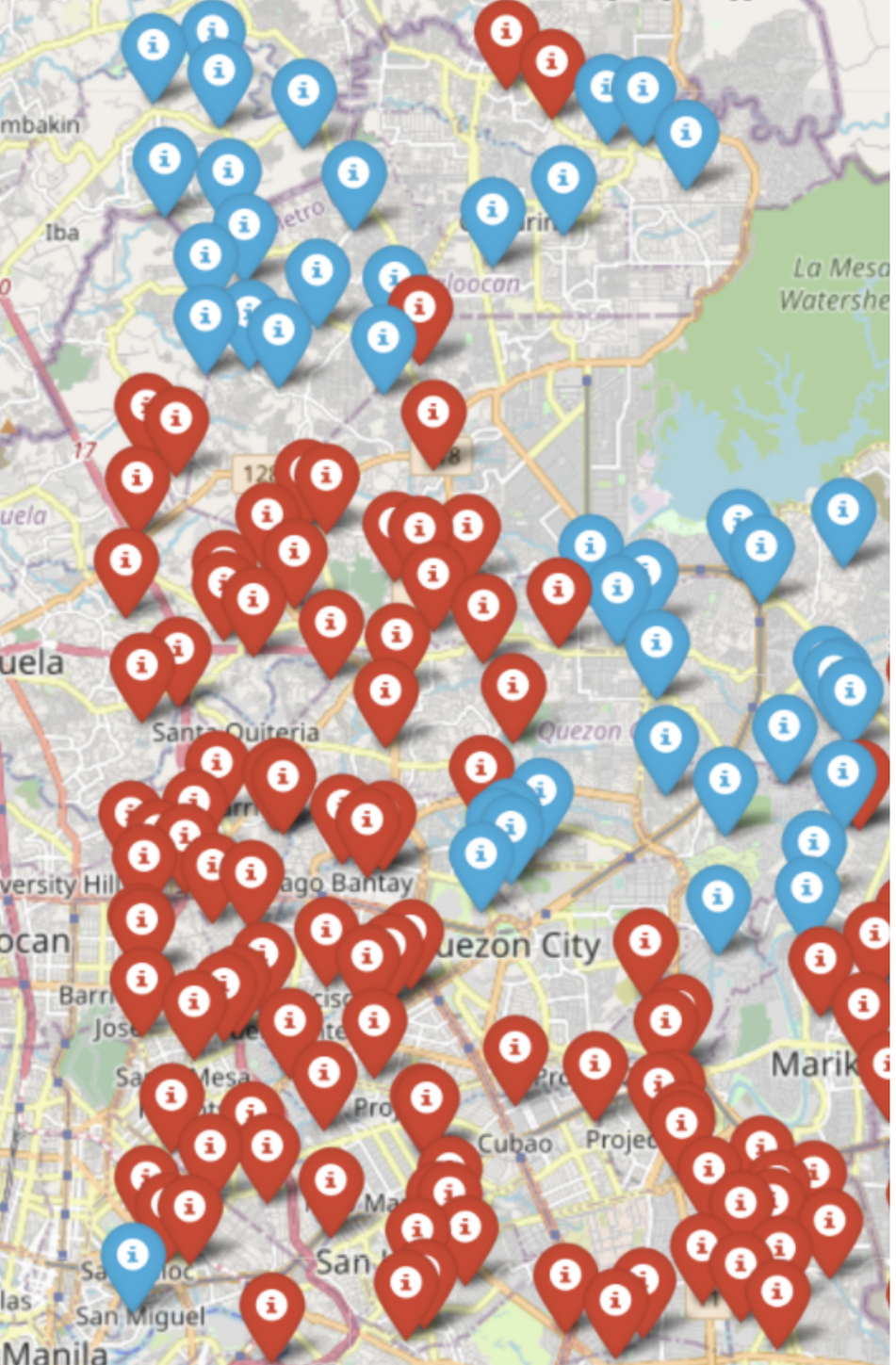


Spatial

longitude
latitude
ISO geo tag (for search)

Temporal

Year
Month
Day



DATA CONSIDERATIONS

CURRENT DATA AUTOMATION

SATELLITE DATA TAKES TIME TO QUERY

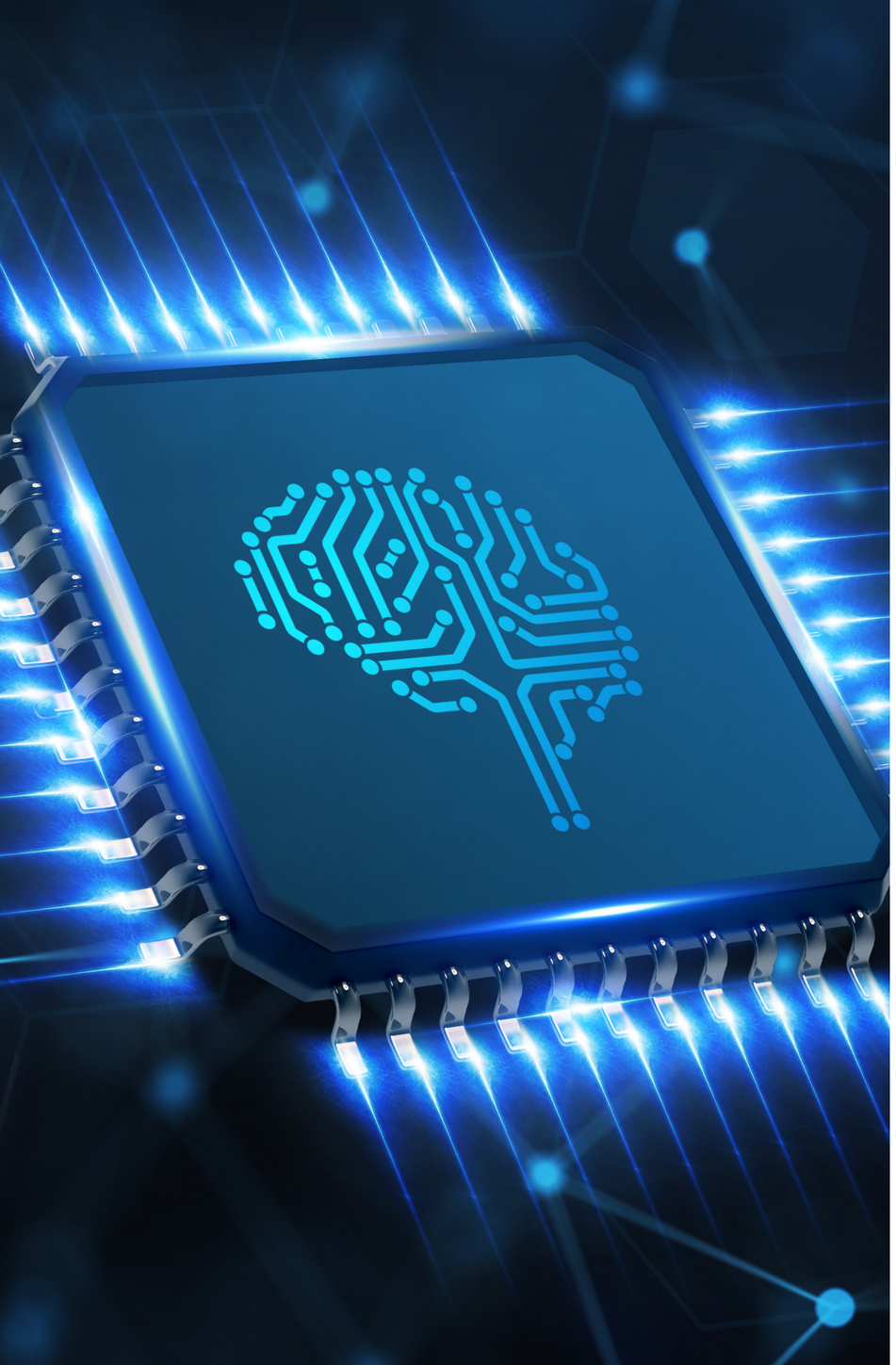
~9 hours for 4.8k data points

OSM DATA IS MEMORY INTENSIVE

Contraction hierarchies crashes a 16 GB RAM machine on 4.8k data points

PYTRENDS LOCATION ISN'T GRANULAR

ISO geo tags are used to limit search to a location



ML MODEL CONSIDERATIONS

CURRENT MODEL SCOPE

TRAINED ON QC, PH DATA

Need data from other locations

EFOI DATA IS IN SCANNED PDF FORMAT

Needs manual manpower to convert to flat file

INFORM RISK MODEL IS USING UNSUPERVISED LEARNING

Risk labels are nice-to-have

DATA BACKLOG

TWITTER DATA

As discussed with BOA, sentiment analysis and NLP are needed to make sense of tweets vs simple engagements/likes/count

NIGHT LIGHTS

Implementation is not straightforward and will be added during future enhancements

FB DATA

BOA has streaming FB data for mobility and will be added when granted access during future enhancements