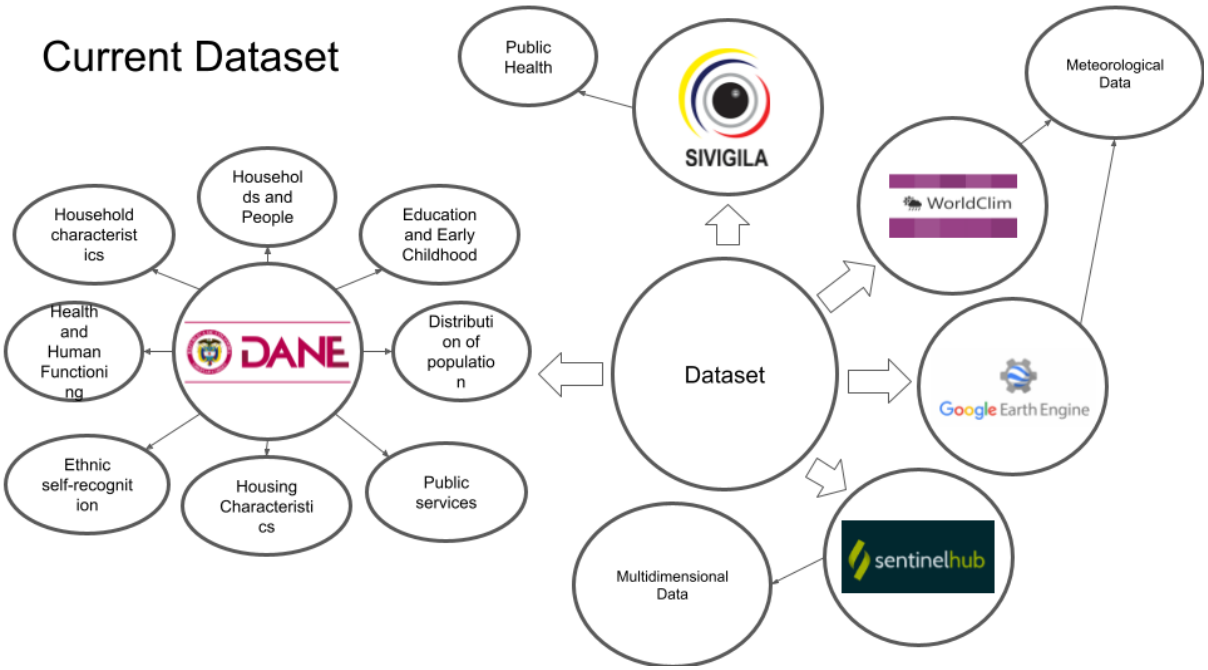
Current Dataset



1- Tabular data (In these section you will find 2 datasets):

1. Dataset with sociodemographic and socioeconomic data (1.2), dengue cases (1.1) and temperature and precipitation (1.3) for all municipalities:

- **dengue_data_all_municipalities.csv**

Municipality code	Municipality	Population/Year	...	Cases/Year	Temperature_month	Precipitation_month	Year/epiweek
5002	Abejorral	20000		23	19	12	12

2. Data set for the Municipality of Medellín that contains refined information on temperature and precipitation (1.4).

- **Dengue_Dataset(Medellin).csv**

cases_medellin	DATA	YEAR	YearWeek	LastDayWeek	MONTH	precipitation_medellin	precipitation_medellin_urbano	precipitation_medellin_rural	temperature
1	2007-w1	2007	200701	16/01/2007	1	12	13	12	19

1.1. Dengue Cases:

(**Data source -> SIVIGILA**):

- Weekly dengue cases based in epi week: From first epi week of 2007 to last epi week of 2019.
- Yearly dengue cases from 2007 to 2019.

Column Name:

- Year/epiweek
    - Example:
        - 2007/w01, 2007/w02, …, 2019/w26

1.2. Sociodemographic and Socioeconomic data:

(**Data source -> DANE)**:

- Municipality unique ID.
    - Column names:
        - Municipality code.
- Municipality name.
    - Column name:
        - Municipality.
- Population for each municipality each year from 2007 to 2019.
    - Column names:
        - Population2007, Population2008, …, Population2019.
- Percentage of the population belonging to a certain age.
    - Column names:
        - Age0-4(%), Age5-14(%), Age15-29(%), Age>30(%)
- Percentage of Afrocolombian Population.
    - Column name
        - AfrocolombianPopulation(%)
- Percentage of Indian Population.
    - Column name
        - IndianPopulation(%)
- Percentage of people with disabilities: This variable describes the group of people who have some physical, psychological or mental limitation.
    - Column name
        - PeoplewithDisabilities(%)
- Percentage of people who cannot read or write
    - Column name
        - Peoplewhocannotreadorwrite(%)
- Percentage of people that have secondary/Higher Education level
    - Column name
        - Secondary/HigherEducation(%)
- Percentage of employed population
    - Column name
        - Employedpopulation(%)
- Percentage of unemployed population
    - Column name
        - Unemployedpopulation(%)
- Percentage of people doing housework
    - Column name
        - Peopledoinghousework(%)
- Percentage of retired people
    - Column name
        - Retiredpeople(%)

- Gender or population in percentage for men and women.
    - Column names:
        - Men(%), Women(%)
- Households without water access.
    - Column name
        - Householdswithoutwateraccess(%)
- Households without internet access.
    - Column name
        - Householdswithoutinternetaccess(%)
- Building stratification.
    - Column names
        - Buildingstratification1(%), Buildingstratification2(%), …, Buildingstratification6(%)
- Number of hospitals per Km2:
    - Column name
        - NumberofhospitalsperKm2
- Number of houses per Km2
    - Column name
        - NumberofhousesperKm2

1.3. Temperature and Precipitation:

c. **Temperature:**
- Temperature monthly for each municipality in Colombia.

d. **Precipitation:**
- Precipitation monthly for each municipality in Colombia.

Column Name:
- VariableName_Month_year
    - Example:
    - PRECIPITATION_jan_07, …, PRECIPITATION_dec_18
    - TEMPERATURE_jan_07, …, TEMPERATURE_dec_18

1.3. Just For Medellin:
- **Dengue_Dataset(Medellin).csv**

Dataset with dengue cases in Medellin, but with weekly temperature and precipitation based on the epidemiological week.

2- Satellite Imagery

- Sentinel-2 weekly images based in epiweek.
- 10 m/px, 12 bands (Nearest neighbor interpolation for bands with less resolution than 10 m/px)
- the best images were obtained using the least amount of clouds algorithm.

In datathon just will be used embeddings of Medellin:
Link: https://github.com/MITCriticalData-Colombia/SatDengue_MakeHealth

Satellite embeddings for 164 images of Medellin can be downloaded here in csv format. The shape for each csv file is given by the structure (164, n_features + 1), where n_features represents the number of features obtained for each model and the extra column is the date of the image:

- **2.1. Satellite images feature extraction with deep learning models**

  - features_resnet50.csv - Download: Feature extraction variation based on resnet50 pretrained on Imagenet - Extracted from Sentinel 2 in Medellin between 2015-2018
  - features_transformer.csv - Download: Feature extraction variation based on vision transformers pretrained on Imagenet - Extracted from Sentinel 2 in Medellin between 2015-2018

- **2.2. Satellite images dimensionality reduction with Variational Autoencoders**

  - embeddings*medellin*100features.csv - Download: Embeddings generated using a variational autoencoder with latent space of 100 (100 features) in csv format - Extracted from Sentinel 2 in Medellin between 2015-2018
  - embeddings*medellin*200features.csv - Download: Embeddings generated using a variational autoencoder with latent space of 200 (200 features) in csv format - Extracted from Sentinel 2 in Medellin between 2015-2018

- **2.3. Satellite images dimensionality reduction with principal component analysis (PCA)**

  - pca*medellin*100features.csv - Download: Embeddings generated using the first 100 principal components in csv format - Extracted from Sentinel 2 in Medellin between 2015-2018
  - pca*medellin*120features(per_band).csv - Download: Embeddings generated using the first 10 principal components in each band (120 features in total per image) in csv format - Extracted from Sentinel 2 in Medellin between 2015-2018