INTRODUCTION

In Houston, TX, we are facing a challenge with flooding in an urban area. To tackle this issue, we are looking to develop a machine learning model capable of predicting the outcome of flood events based on their initial conditions. We have a collection of 3,000 simulated flood incidents (training set), each serving as a historical data point to train our model.

The assignment's goal is to build a data-driven model to capture the simulator's logic that generated these incidents and to replicate its predictions. Each simulated incident is unique, starting with different initial conditions, leading to varied outcomes. However, the geographical layout, including the street network and the location coordinates of each street segment, remains unchanged throughout all simulations. This consistent geographic data is available in the **'edge_info.csv'** file.

Additionally, specific parameters that influence how the flood unfolds in each simulation are set at the beginning and differ from one incident to another. These parameters are detailed in the **'training_parameters.csv'** and **'test_parameters.csv'** files for the training and testing datasets, respectively.

Some basic definitions:

- **Nodes**: Intersections or endpoints of streets. Each node has a unique *9-digit* identifier. Like: **152356047**
- **Edges**: Street segments linking two nodes, defined by **'head_id'** and **'tail_id'**.

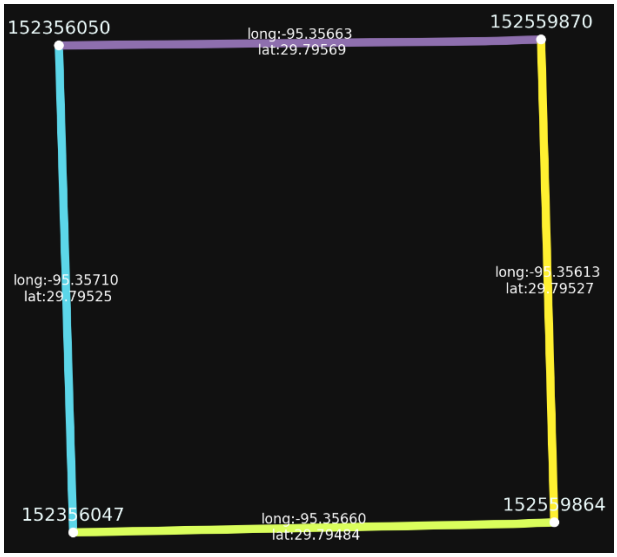

Figure 1 – Four street segments (edges)

Figure 1 shows an isolated block consisting of 4 street segments. The cyan-colored edge (in the left) can be represented by its head and tail nodes: **152356050**, and **152356047**, respectively.

The selected urban area is composed of 191 edges that can be shown in Figure 2:

Figure 2 – The edges of the urban area

DATASET OVERVIEW:

You received 3,000 observations, each detailing the initial and final states of separate observations. These observations represent flood progression in a hypothetical urban layout, determined by street connectivity, elevation, and infrastructure.

'./training_parameters.csv'

SurfaceType	Waterflow	InitialWaterLevels	DrainageSystemCapacity	GreenSpaceRatio	ObservationIndex
C	4	3	0.14	0.21	0
C	5	5	0.26	0.19	1
B	2	3	0.17	0.15	2
D	5	5	0.14	0.14	3
D	3	4	0.2	0.23	4
D	2	5	0.12	0.13	5
A	5	4	0.16	0.2	6
B	3	4	0.2	0.17	7
C	4	3	0.28	0.12	8
B	2	3	0.26	0.26	9

'./training'

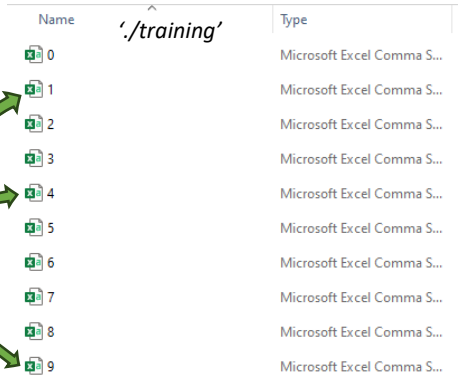

Figure 3 – initial and final states of each observation can be found in a separate csv file in 'training' folder. These CSV names are consistent with the ObservationIndex column in 'training_parameters.csv'

Simulation parameters for each observation are retrievable from `./training_parameters.csv` and `./test_parameters.csv`. These parameters include:

- **ObservationIndex**: an identifier for the observation.
- **SurfaceType**: Type of urban surface.
- **Waterflow**: (it's named **RainfallIntensity** in the CSV file) Intensity and duration of water flow.
- **InitialWaterLevels**: (it's named as **Init_Max_hour** in the CSV file) Pre-simulation underground water level.
- **DrainageSystemCapacity**: Indicator of drainage efficiency.
- **GreenSpaceRatio**: Proportion of greenery in the urban area.

To make it simpler, the above-mentioned parameters are the average of those parameters in the urban area. We don't have the microscopic values for each edge/street.

Within the `./training` and `./test` directories, CSV files correspond to the observations. Each file, named after the ***observation index***, records various edges with attributes:

- **head_id**: ID of the head node.
- **tail_id**: ID of the tail node.
- **flooded_init**: Whether the edge was initially flooded.
- **flooded_final**: Whether the edge was flooded after the simulation (this is the target variable!).

Extra information about the edges can be found inside `edge_info.csv`:

- **longitude**: Longitude of the edge's center.
- **latitude**: Latitude of the edge's center.
- **altitude:** the elevation of the edge's center.

.

Figure 4 provides a visualization of some observations (at their final state) where the grey edges are non-flooded edges, and the purple (**flooded_init**) together with cyan ones are the **flooded_final** edges in each observation:

Figure 4 – Visualization of 4 observations – purple edges are the initial flooded ones; cyan edges are the ones that became flooded after the initial ones.

Bear in mind that this visualization is just for you to understand the logic behind it. You're not asked to do image processing. However, if you'd like to do that, you can convert the files to PNG images, using `visualize` function that is given in the Colab notebook.

OBJECTIVE:

Your task is to devise a predictive model that uses this data to predict the **flooded_final** column for each sample in the test set (`./test`).

This task is crafted to provide insight into handling practical machine learning challenges, where an ideal solution is not always attainable. Do not worry if your model doesn't perform as well as your peers'—the emphasis here isn't on achieving the best prediction results. Instead, we value the originality and soundness of the methods you employ in your approach.

DELIVERABLES:

Submit a ZIP file named with your student id (like `10621111.zip` – if you're a group: `10621111-10621112.zip`, use dashes to separate your student ids). As all the files will be automatically compiled, please make sure of the naming.

The ML Assignment – 2023-24

The zip file contains:

- Your team information (your student ids, first name, surname) in a .txt file, named: **`info.txt`**
- The trained model (serialize it using Python's **pickle** module).
- Your Colab/Jupyter Notebook (**.ipynb** file), with comprehensive documentation of your methodology.
- Predictions for the test set, in individual CSV files mirroring the training set's structure.

EVALUATION CRITERIA:

Submission Deadline: **January 6th, 2024** - at **23:59**

Submissions will be assessed based on:

- Prediction quality on the test set.
- Originality and creativity of your approach.
- Clarity and detail in your analytical documentation.

I'm looking forward to your innovative solutions to this practical and impactful problem.

Regards,

Masoud Jalayer, PhD