# COMP 6940: Big Data and Visual Analytics - Assignment 4

University of the West Indies, St. Augustine

Due Date: April 19th @ 11:59 PM

## 1 Time Series Analysis (50 marks)

The datasets for this section can be downloaded here.

### 1.1 Forecasting

1. There are two datasets. You are required to merge them into a single dataframe such that we have $CO_2$ emissions and the terawatt hours consumed by each country over the years.
2. Ensure that the data is 'cleaned' before proceeding further.
3. Select 2 countries to investigate trends in emissions. Explain briefly the reason for your choices.
4. Identify any trends in annual $CO_2$ emissions from oil (per capita) over the years. Compare the trends for two countries.
5. For the two selected countries, forecast using simple exponential smoothing the expected "$CO_2$ emissions from oil (per capita)" for the most recent five years, use the RMS metric to indicate the accuracy of your forecast.
   (a) Choose the alpha that results in the lowest RMS.
   (b) Plot a graph showing the actual value and the forecasted values.
6. Repeat Question 4 for 'Coal Consumption - TWh' and another column of your choice. Explain why you chose those two columns.

## 2 GPU-accelerated Data Analysis with RAPIDS (50 marks)

### 2.1 Setup

Ensure you have access to a GPU-enabled environment in Google Colab and have successfully installed the RAPIDS suite as detailed in the preparatory materials. Download the NYC Taxi Trip Duration dataset from Kaggle or from here.

### 2.2 GPU-Accelerated Mchine Learning

In this section you will do a practical application of GPU-accelerated data analysis using the RAPIDS suite of libraries. The focus will be on leveraging the New York City Taxi Trip Duration dataset from Kaggle, applying Gradient Boosting Machine (GBM) models for predictive analysis, with a particular emphasis on comparing the speed and efficiency of GPU-accelerated computations with traditional CPU-based methods.

1. **Comparative Data Processing**
   (a) Perform data loading and preprocessing tasks first using pandas (CPU) and then replicate the same tasks using cuDF (GPU). Document the time taken for each operation in both scenarios.
   (b) Conduct basic exploratory data analysis (EDA) with both CPU-based tools (e.g., matplotlib) and GPU-accelerated tools, noting any differences in performance and responsiveness.
2. **Feature Engineering and Selection**

(a) Engage in feature engineering, creating new variables that could aid in predicting trip durations. Compare the execution time for these operations on CPU vs. GPU.

(b) Select relevant features for the model based on their correlation with the target variable, assessing the speed of these operations on CPU and GPU.

3. **Model Training and Evaluation**

   (a) Train a Gradient Boosting Machine (GBM) model on the dataset using scikit-learn (CPU) and cuML (GPU). Record and compare the training times.

   (b) Evaluate the accuracy of both models and document the time taken for predictions on the test set using CPU and GPU.

4. **Performance Analysis**

   (a) Compile and compare the execution times for tasks performed on CPU vs. GPU, creating a detailed analysis of the observed performance differences.

   (b) Reflect on the implications of these findings for data science workflows, particularly in terms of efficiency and scalability.

# 3 Deliverables

Students are required to submit the following:

1. A detailed Python notebook containing all code written and comments documenting the process for part 1 and part 2 named **a4_part1_idnumber.ipynb** and **a4_part2_idnumber.ipynb** respectively.

2. Zip all files into a single file called **a4_idnumber.zip**.

3. Do **not** submit any dataset files.

4. See the Course Github Page for further submission details.