# Big Data and Intelligent Analytics
## Spring Semester 2022

INSTRUCTOR: Sri Krishnamurthy
analyticsneu@gmail.com

## Assignment 1: Experiments with Big data

In this assignment, you will work with large datasets to ingest, process, store it so you can access it through different means.

**Preparation:**

1. https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/examples/SEVIR_Tutorial.ipynb
2. https://www.ncdc.noaa.gov/stormevents/ftp.jsp
3. https://cloud.google.com/bigquery/docs/visualize-data-studio
4. https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Paper.pdf
5. https://sevir.mit.edu/sites/default/files/About_SEVIR.pdf

**Case:**

**You are a freshly minted data scientist engineer at WeSpace Inc. which leverages space and weather data to build forecasting systems! It is exciting and you just bought your first Tesla to celebrate. Post Covid, you drive to your office and your manager is kind enough to offer the corner office with the view of the ocean on one side and mountains on the other. She checks in to make sure you don't miss the freshly baked cookies that are at the kitchen. During group lunch, the team is engrossed in a fresh challenge. They plan to build a nowcasting system leveraging satellite data and weather datasets but everyone in the team wants to build models!**

**Where is the data ? You ask. Everybody stares at you and pretend it isn't a problem and continues the conversation about Deep Learning and other problems. Your manager interjects and asks your question to the group again. "That is an important question! Where is the data?".**

**Some team members say it is all on AWS and in csv files and HDF5 format.**

**"How do we plan to access it?" You ask!**

**"Well, we will get there when we get there!", a team member says!**

**Your manager in your 1:1, commends you for asking the hard questions and gives you your first assignment! First, we need to try out how to access the data. She suggests 3 architectures and asks you to experiment and provide your recommendations.**

**----------**

**Tasks:**

1. **Review the links above to prepare for the assignment**

## Working with Large datasets

2. **Review the following links**
    a. **https://github.com/MIT-AI-Accelerator/eie-sevir/blob/master/CATALOG.csv**
    b. **https://www.ncdc.noaa.gov/stormevents/ftp.jsp**
    c. **https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/Storm-Data-Export-Format.pdf**
3. **Download the data flies from SEVIR and from the Stormevents ftp site for Event ID: 835047. Note: For downloading specific files, you can use both.. see https://github.com/MIT-AI-Accelerator/sevir_challenges**
4. **for examples**
5. **Modify** https://nbviewer.jupyter.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/exam

ples/SEVIR_Tutorial.ipynb and redo the tutorial for files corresponding to **Event ID: 835047**

**Analyzing the metadata in Datastudio**

6. Download the **Catalog.csv** file and the **storm event files** corresponding to the date range in the Catalog.csv file.
7. Use the skills learnt in the Google tutorial(https://cloud.google.com/bigquery/docs/visualize-data-studio ) to build a dashboard that would analyze the storm event files and the Catalog.csv files

**Deliverables (Due Feb 11th 11.59am):**

1. A 2-5 page report in https://github.com/googlecodelabs/tools format to illustrate your understanding of various steps and outcomes.
2. Github with
    a. Links to the notebook and any other supporting files
    b. Links to dashboard
3. You will be given 10 minutes to present your company analysis in class on Feb 11th