

# Big Data and Intelligent Analytics

## Spring Semester 2022

INSTRUCTOR: Sri Krishnamurthy  
[analyticsneu@gmail.com](mailto:analyticsneu@gmail.com)

---

### Assignment 2: Designing a Nowcasting Intelligent System

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

#### Preparation:

1. Review:
  - <https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Paper.pdf>
  - <https://papers.nips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Supplemental.pdf>
2. [https://sevir.mit.edu/sites/default/files/About\\_SEVIR.pdf](https://sevir.mit.edu/sites/default/files/About_SEVIR.pdf)
3. <https://github.com/MIT-AI-Accelerator/neurips-2020-sevir>

#### Case:

Now that you demonstrated your understanding of the SEVIR datasets in Assignment 1, your team wants to try out different models that can leverage these datasets. The Data science team has told them that they will be building models using machine learning and your team (data engineering) will be expected to help automate and productionize the analytics. They came across the papers[1,2] listed above and want to replicate the experiments. The team is particularly interested in temporal datasets and in synthetic data generation use cases and the paper provides them exactly those examples : Nowcasting (temporal datasets) and SynRad generation (cross sectional data). Your team wants to familiarize yourselves with the models (3) and asks you to replicate the models.

---

-----

**Tasks:**

**Part 1:**

1. Get familiar with the models by reading the papers[1]
2. You are not expected to train the models. But you will work with pre-trained models. Download the models (<https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/models> )
3. Start with the Synrad usecase in [3] and run the notebooks in <https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/notebooks>
4. Try out the sample datasets and complete the notebooks
5. Note that the sample datasets for the Nowcasting notebooks are missing. You will need to create test datasets to try out the models.
6. Generate test data using the generators in <https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/src/data>
7. Start with Synrad to get a feel for the generators and generate data for 2-3 eventids. Compare the formats you generated to the ones given in the sample datasets. You can use <https://www.hdfgroup.org/downloads/hdfview/> to view the files.
8. Now create datasets for the Nowcasting usecase and try out the sample notebooks <https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/notebooks>

---

## Part 2: Designing a Nowcasting Intelligence System

After you demonstrate the model, the project sponsor is impressed. They got the budget to deploy the model so that they can sell the Nowcasting Intelligence system. They ask you to put together a functional requirements document for the Nowcasting intelligence system. The requirements should cover the details discussed in class and summarized below.

1. Each team will focus on one use case.
  - 1.1. Government - US (for Forecasting)
  - 1.2. Airplane - Air Force
  - 1.3. Maritime Use - Ships
  - 1.4. NOAA
  - 1.5. Media Channels/Outlets
  - 1.6. National Grid or Equivalent Electricity Supply Corporation
  - 1.7. Farmers/Agricultural Industry
  - 1.8. Insurance
  - 1.9. Consumers - WebApps(like Weather.com)
2. The system should be able to accomplish two goals:
  - 2.1. Given 13 images of cloud cover -> Forecast 12 images
  - 2.2. Backtest the model for 10-20 use cases using data from different event ids
3. The system will be distributed as follows:
  - 3.1. Python Package (distributed python package) (See <https://packaging.python.org/en/latest/tutorials/packaging-projects/> )
  - 3.2. WebApp (See <https://streamlit.io>)
  - 3.3. API access(model-as-a-service) – REST API (See <https://fastapi.tiangolo.com>)
4. The architecture for the system should be discussed (use <https://diagrams.mingrammer.com> and see <https://systeminterview.com> for examples). It should cover:
  - 4.1. Model
  - 4.2. Data - Sevir - AWS - S3 bucket
  - 4.3. Data storage for inputs/outputs
  - 4.4. Hosting options
  - 4.5. Compute(On premises or rent on a cloud)
  - 4.6. Visualization Tools(Like Power BI/Tableau)
  - 4.7. OS(Command line Shell) to interface with the models
5. You should discuss how you plan to distribute the package
  - 5.1. How will you document it? (Google codelabs? <https://www.sphinx-doc.org/en/master/> ? <https://www.mkdocs.org/user-guide/writing-your-docs/> ?)
  - 5.2. How will you distribute it? (pypi ? conda? Containers with Docker?)
  - 5.3. How will you advertise the APIs? (examples, playground?)
6. Develop a skeletal user interface using Figma for the use cases you plan to accomplish

---

**Deliverables (Due Feb 25th 11.59am):**

1. A 2-5 page report in <https://github.com/googlecodelabs/tools> format to illustrate your understanding of various steps and outcomes for Part 1
2. Google Codelabs that details the functional requirements doc for Part 2
3. Github with
  - a. Links to the notebook and any other supporting files (Diagram etc)
4. You will be given 10 minutes to present your analysis in class on Feb 25th