Cyberinfrastructure Competition

Welcome to Cyberinfrastructure Competition!

Cyberinfrastructure involves computing systems, data, software, visualization, and people. It is widely used for data-driven research and scientific discovery. The goal of this cyberinfrastructure competition is for the participants to use various components in cyberinfrastructure to solve an open-ended research problem related to large-scale spatial data in an interdisciplinary team. The participants will go through the entire process of finding/downloading relevant data, analyzing data, to visualizing and presenting the results. Through this process, the participants will demonstrate their skills of using cyberinfrastructure and interdisciplinary collaboration.

The participants will be divided into teams. Each team has 2-3 students covering computer science background and geoscience background (please see the teams later). Each team has 2 weeks, from 1/19/2024, Friday to 2/2/2024, to solve the problem described below.

Pre-workshop Survey

Please fill in pre-workshop survey at (link removed)

Problem Description

The project involves two aspects, one is a machine-learning based prediction and the other is geospatial analysis. The dataset (including two files, as described below) can be downloaded from

https://www.dropbox.com/scl/fo/rif6qta0cqgfp3x4hkd7l/h?rlkey=shdxlb548yweolfcsg2lg6034&dl=0

Machine-learning based Prediction

You will be asked to design a spatio-temporal machine learning model to learn and predict the out and in flows at the Manhattan Island in New York City (NYC). Basically, imagine you are a New Yorker and you want to ride a bike from one station A in Brooklyn and arrive at another station B in Manhattan. Then a bike trip is formed by a start bike station with the bike pick-up timestamp followed by an end station with the drop-off timestamp.

The participants will be given the following files for the spatio-temporal machine learning model training.

- RawData.csv: contains the raw data of the bike sharing stations, trips, and rider information for the month of October 2019.
- TraininingData.h5: contains the processed and formatted spatio-temporal tensor obtained from RawData.csv. Basically, the entire geographic region studied is [minimum latitude, maximum latitude, minimum longitude, maximum latitude] = [40.69331716, 40.76915505, -74.01713445, -73.95381995]. We have partitioned this region into a total

of H x W grid cells, where H and W respectively represent the numbers of rows and columns if you imagine the entire region as a two-dimension image. The spatio-temporal tensor **F** is of the dimension T x 2 x H x W, where T represents the total number of time slots (30min for each slot, starting from 2019/10/01 00:00:00 till 2019/11/01 00:00:00), H and W are the number of rows and columns of the grid cells in each time slot, and 2 represents the out and in flows. For example, **F**[1,0,0,2] = 12 means a total of 12 bikes have been picked at the time slot 1 at the row 0 (vertically) and column 2 (horizontally) in the partitioned map. Similarly, **F**[1,1,0,2] = 14 means a total of 14 bikes have been returned at the time slot 1 at the row 0 (vertically) and column 2 (horizontally) in the partitioned map. Another variable in the TrainingData.h5 contains all the T timeslots in the training data that corresponds to the calculated out and in flows. For instance, 2019100100 represents the first 30min time slot (00:00 – 00:30) of October 1, 2019, where yellow part represents the day (10/1) and the green part represents the index of the time.

Your task is develop a machine learning based flow prediction algorithm (i.e., prediction the number of pick-ups and returns in each grid). Specifically, the problem is as follows:

Given the spatio-temporal **F** that is of the dimension T x 2 x H x W for model training, the participants need to develop a machine learning model that can predict the elements of a spatio-temporal tensor **F'** (in T' x 2 x H x W) that belongs to a different time period.

Training and validation: you can partition the tensor data based on the first dimension (i.e., time) into two separate periods, one for model training and the other for validation. We will hold out the testing data that is different from the training data, and evaluate the submitted trained model as well as the code (more details on submission will be released later on).

For testing, you will be given 5 weeks of data, and the evaluation is based on the prediction accuracy of the last four weeks (in case you plan to use the first week as the input for an end-to-end machine learning model). The performance metric is the mean square error (MSE).

To help you work on the project, we will release a basic tutorial (on Monday, 1/22/2024), which you can use as a reference. Note that this tutorial is very basic – it is meant to explain the loading and pre-processing of the dataset. It requires substantial improvement in data analytics to obtain better prediction results.

Geoscience analysis. The participants can geocode the raw data using ArcGIS. Examples of potential directions of GIS spatial analysis are listed below. These are just some examples. Feel free to conduct any spatial analysis beyond these examples (you can use the attached paper for more ideas).

- Spatial Analysis of Usage Patterns: By mapping trip start and end points, GIS can reveal spatial patterns in bike usage. This helps in understanding which areas have the highest demand for bikes and which routes are most popular.
- Temporal Analysis: Combining GIS with time data, you can analyze how bike usage varies throughout the day, week, or year. This helps in identifying peak usage times and seasons, which is crucial for efficient fleet management and maintenance scheduling.

- Network Analysis: GIS can be used to analyze the connectivity and efficiency of the bike-sharing network. This includes identifying gaps in the network, optimizing the location of bike docks, and improving accessibility to the service.
- Integration with Other Transportation Modes: GIS analysis can show how bike-sharing systems integrate with other modes of transportation, like public transit. This can help in planning combined mobility solutions.
- Accessibility Analysis: Identify areas with limited access to bike-sharing services, which can inform decisions on where to expand the network.
- Demographic Analysis: By overlaying bike-sharing data with demographic information from the census, you can understand who is using the bike-sharing system. This includes insights into age groups, income levels, ethnic backgrounds, and other demographic factors. You can search and download US 2020 Census data: Just search for "current year ACS" from ArcGIS Pro: https://www.arcgis.com/apps/Cascade/index.html?appid=c6a35a434a8f4913b9c35022290efdd0

Judging Criteria

Each team present your results in a final presentation, and answer questions. The presentation will be on Friday, 2/2/2024 (the exact time to be determined). A committee of judges will rank the teams based on your presentation using two main criteria: technical merit and interdisciplinary collaboration.

Your presentation should include the following components to showcase the Technical merit and your interdisciplinary collaboration:

- All team members need to participate in the presentation.
- The presentation starts with introduction (on your process in approaching the problem), approaches, data set used (in addition to the dataset that is provided to you), results, and reflection.
- Each team member needs to present (part of) the materials that are outside your background (i.e., the work of your team member). Clearly mark this in your presentation. The more you present about others' work, the better.
- Approaches: what approaches you used in the fields of computer science and geoscience. Both aspects are required.
- Results: results from both computer science and geoscience aspects.
- Reflection: Describe one experience that interdisciplinary collaboration works for your team? What do you think happened?
- Reflection: Describe one experience that interdisciplinary collaboration did not work for your team? What do you think happened?

  More detailed judging rubrics will be released later in the competition.

Coaching Sessions

The organizers will help you during the competition. In particular, each team has two coaching sessions with the organizers: one in the first week and the other in the second week, each 10-20 minutes. Beyond these two sessions, you can feel free make an appointment with the organizers under special conditions (e.g., you are stuck, conflict between team members, etc.)

- Coaching session in the first week: This is to answer your questions regarding the problem and provide some suggestions on your approach. You can describe your thoughts and plans of solving the problem. Schedule your session by Wednesday of the first week.
- Coaching session in the first week: This is primarily to show your results so far and seek suggestions. Schedule your session by Wednesday of the second week.

Schedule

- Day 1, Friday 1/19: competition released.
- Day 2-14: work on the project, and prepare for presentation.
- Day 14 Friday 2/2: team presentation.
- 2/9 Friday: announcing awards.

Make sure that you give yourself enough time to work on the presentation. If some of your results are not exactly satisfactory to you, that is fine – there are two criteria in judging the outcome of your project.

Teams

This competition contains the following teams, which are formed by the organizers based on the participants' background.

Teams (names removed)

A few tips on team/interdisciplinary collaboration

We hope that you will enjoy the teamwork and interdisciplinary collaboration, which is often needed in work place, whether you are in industry or academia. The work involving cyberinfrastructure tends to particularly involve collaboration among team members with diverse backgrounds. Some tips:

- Getting to knowing and brainstorming with each other. Within each team, it is a good idea to start with a brainstorming session. You can introduce each other, getting familiar with each other's backgrounds. Take a look at the problem together and make some decisions on how to proceed.
- Meet regularly. Meet again soon after the first meeting, and then meet regularly. In-person meetings are preferred, particularly at the beginning.
- Some basic ways of working together: take turns to talk and describe your thoughts. Keep in mind that the team members have very different backgrounds (which was intended), and hence you may think very differently and have very different skill sets. You want to use those as an advantage. Keep in mind that you want to respect each other's opinion, despite the circumstances.

- Avoid a complete “divide-and-conquer” style. This can naturally happen in an interdisciplinary team. While it can sometimes make your work more efficient, it can also lose the strength of your interdisciplinary team. Frequent meetings and keeping yourself open-minded can help avoiding such issues.

Awards

First place: $300. Second place: $200. Third place: $100. Ties in places are possible.