CAPSTONE PROJECT PROPOSAL | Data Science Institute at Columbia University

A project proposal submission consists of a project description and a dataset. It can be about a novel dataset or a publicly available dataset with a novel task.

Email address *

wangchenxi@didiglobal.com

Please read BEFORE you propose a project

Unless otherwise noted (e.g., public data sets), students are instructed not to share the capstone dataset outside the team or use the data beyond the scope of the capstone project. Please propose a project with the expectation that its results will be shared publicly. Capstone projects are an important educational experience for our students and also provides them with a showcase to share in interviews for potential employers or for various other audiences inside and outside Columbia.

Columbia views capstone course work as work owned by the students and shared with the mentors. As a general practice to simplify the capstone project program, we do not sign NDAs for the capstone projects.

1. DSI INDUSTRY AFFILIATE *

Please select your company.

DiDi ▼

Project Description

2. MOTIVATION, BACKGROUND AND OVERVIEW *

Please state briefly what is the problem that the project tackles. The projects need to be focused on a data science problem that is engaging, relevant, clearly defined and of the right scope for a semester. When assessing the proposals we will be looking for a diverse set of problems that address different topics and technical requirements that our students can address. The evaluation criteria will include: Is this a data-science project? Can our students learn about a data science application in the real world? Is the proposed research problem important and can potentially have a big impact? Will our students be excited about it? Please provide your project description having these criteria in mind.

This project is about analyzing the relationship between work behavior of drivers on a ride-hailing platform and their in-service hours. The in-serice hours of a driver is the accumulated length of time within which the driver is serving an order. It is closely related to his/her income, thus an important metric to monitor.

On a ride-hailing platform, drivers are free to choose when to get online, how long to stay online, and where they go when they are not assigned an order, typically based on their habits and preferences. On the other hand, it is apparent that even for drivers working over similar length of time, their in-serivce hours could show significant differences. This project tries to tackle the problem of understanding the relationship between driver in-service hours disparity and their work patterns, if any. The insights obtained from this task would be very useful to the platform to better serve the drivers' need as well as to improve the efficiency of the platform.

The students will be given a large trip and trajectory data set through DiDi's GAIA program. It is open for them to use the data science technique of their choice to approach this problem, e.g. prediction factor analysis, inverse reinforcement learning.

DATASETS

The dataset(s) can be public or private. Please keep in mind that the students will need to list the project on their CV and the report will be public. All datasets must be submitted by Friday, July 19, 2019 for Fall 2019.

3. DATASET *

Please provide a detailed description of the type of data that is required to address the problem. For example, is this social media data, medical data, financial data, etc? What is the size of the data. Will the organization provide the majority of the data or is the data accessible via other avenues/ sources? How much of the data is available? Do the students need to gather data? In assessing the projects, the availability and type of data will play an important role. Please consider these evaluation criteria for data requirements when submitting the proposal: Is the data set clearly defined? Is the data set complex and big enough for creating learning opportunities? Is the data set ready? (availability, need for processing) Does the data require extensive computing resources (if yes, can the affiliates provide resource/funding?)

GAIA Open Dataset, https://outreach.didichuxing.com/research/opendata/

4. Data Type *

Public data is data made available by a third party and is available to the general public. Novel data is data that has been recently published by the proposer or will be made public as part of this project. Private data is data that can not be made available after the project ended. Please check all that apply.

□ Uses Public Data

□ Uses Private Data

☑ Uses Novel data

5. How will the datset be made available? *

For example: CSV/XLS file, remote database, raw images or documents, REST endpoint, etc.

https://outreach.didichuxing.com/research/opendata/

GOALS, OUTCOME and SKILLS

6. RESEARCH GOALS *

What is the goal of this project? What questions do you want answered? What has been done already to achieve this goal?

The goal of this project is to investigate the factors (as indicated by the trip and trajectory data) contribute to the differences in driver in-service times.
Questions to be answered:
What differentiates an experienced driver on a ride-hailing platform from novices?
Are there any particular work habits that lead to gaps in driver in-service times? If so, what are they?

What are the ethical considerations?

Are there any ethical concerns about the proposed project such as privacy, transparency, and bias that we should pay special attention to?

No

What is the relevant background needed for the project?

In order to make sure we build the right team of students for each project, please provide information on the relevant background information that someone working on the project should have. What technical skills they should have and/or relevant literature (please provide citations) or tools (please provide links) they will need to know or able to learn.

machine learning, statistical analysis, programming in Python, familiar with machine learning packages, e.g. scikit-learn, Tensorflow

OUTCOME *

What deliverable do you expect from this project?

☑ Model

☑ Report

☑ Paper

☑ Software

□ Other: ____________________

SKILLS *

What skills should students expect to learn through their project? Check all that apply.

☑ Project planning and scoping

□ Data acquisition and scraping

□ Data versioning and management

☑ Data cleaning

☑ Combining data sources

☑ Exploratory data analysis and visualization

☑ Supervised modeling

□ Unsupervised modeling

□ Establishing evaluation metrics

□ Working with text data

□ Working with image data

☑ Working with time series data

□ Working with tabular data

☑ Working with geospatial data

□ Other: ____________________

What are the quantitative and/or qualitative metrics that can be used to judge the successful completion of the capstone project?

Solid ideas, experiment and implementation

Are international students on a F1 or J1 student visa eligible to work on this project? *

○ Yes

● No

PROJECT MENTORS

An important aspect of the capstone project is the opportunity for students to work with professionals across different industries or academic research labs. Thus each organization must provide mentorship to the students so that they can receive constant feedback and guidance (while each team will also have a faculty advisor the organization mentor will play a crucial role in guiding the team). Please specify who will work with the students and what are their qualifications or training? What amount of time per week do they intend to devote to working with the project team? Each mentor will also ideally help the DSI faculty advisor assess the success of the project at the end which will translate to a grade for the students.

Each capstone project will be mentored by at least one industry mentor and one faculty mentor, with the industry project proposer(s)/mentor(s) as the primary mentor(s). Industry mentor(s), in addition to monitoring the project progress and provide timely guidance to the capstone team, are expected to

1. Meet with the team on a bi-weekly basis (teleconference is fine)
2. Review the midterm progress report and provide comments to both the team and the course instructor
3. Attend the final poster presentation session
4. Review and evaluate the team's final report
5. Provide comments on each team member's participation

Mentor 1

Name *

Tony Qin

Title *

Head of RL Group

Department/Division

DiDi Labs

Email Address *

qinzhiwei@didiglobal.com

Phone Number *

646 3000486

URL to LinkedIn public profile [optional]

Current Resume - Upload [optional]

What amount of time per week do you intend to devote to working with the project team?

1-2 hours

Mentor #2 (Optional)

Name

Chenxi Wang

Title

Senior Research Program Manager

Department/Division

DiDi Research Outreach

Email Address

wangchenxi@didiglobal.com

Phone Number

N/A

URL to LinkedIn public profile [optional]

Mentor #2: Current Resume - Upload (Optional)

What amount of time per week do you intend to devote to working with the project team?

1 - 2 hours

16. Are you willing to work with two teams of students? *

● Yes

○ No

17. If yes, please indicate how the project may be appropriate to engage two teams.

Focus on one challenge with different approaches