DataFirst Handbook

Last Updated October 4, 2023

Background

The USC Center for Knowledge-Powered Interdisciplinary Data Science (CKIDS) provides opportunities to get involved in collaborative data science projects with other faculty and students across the university and with data science students in training.

In the past years, faculty members across several schools at USC have been collaborating through CKIDS to work on joint projects through DataFest events. In these events, faculty and senior researchers have been able to tackle new interdisciplinary topics, and engage students in data science, computer science and other disciplines to work together to formulate interesting problems and to define joint approaches to solve them.

Any USC faculty, post-docs, or senior researchers interested in exploring new collaborations in data science can propose projects to the center. A call for project proposals is issued every semester.

Faculty

The projects proposed should be semester-long projects where each student spends around 8-10 hours a week.

Students may contact you directly throughout the application process. We aim to minimize the email burden on advisors, but if you find direct emails from students to be useful that is fine. After the assignments to projects are completed, you may receive email appeals from students indicating interest in your project, but these are often mass mailings.

Students

Faculty will expect that students spend around 8-10 hours a week. Students will

- Project Mentor(s) should meet with students once a week
- Project Advisor(s) should meet with students and project mentor once a month
- DataFirst Chair(s) can meet with students once in the semester if requested

Timeline for the Semester

Once the assignments for students to projects and mentors, the timeline would start:

Timeline	Report Type	Length	Sections
Week 2	**Data Assessment Document**	2-3 pages	1. Data overview and examples 2. Data accessibility (eg, files, databases) 3. Data formats 4. Data challenges (eg, heterogeneity, size, pre- or post- processing needs, etc)
Week 3 or 4	**Requirements Document**	Up to 5 pages	1. Motivation 2. Problem for the semester 3. State of the art (summary) 4. Design and general approach 5. Use case scenario 6. Desired outcomes and benefits
Week 6	**Mid-Project Presentation**	7 slides	1. Problem overview 2. Use case example 3. Desired outcomes and benefits 4. Approach 5. Results to date 6. Planned work for the remainder of the semester
Week 10	**Final Project Report**	Up to 10 pages	1. Final Requirements Document 2. Final Data Assessment Document 3. Approach 4. Results to date 5. Products (eg new software, new data, etc) 6. Future work
Week 11	**Final Presentation**	6 slides	1. Problem overview 2. Approach 3. Use case example 4. Benefits of outcomes 5. Results 6. Future work

Data Assessment

Introduction

This page serves as a Data Assessment Document for the project. It should be no more than 2-3 pages long. It can be drafted in the first two weeks of the project, and can serve as an interim project report. It can be refined as the project progresses and the use of the data is better understood.

Data Overview and Examples

Give a brief description of the data provided for this project, what it represents, how it was collected, and why it may help address the problem you are tackling. Discuss if you will be using all the data or only some subset of it for the project. Consider possible additional data that may be publicly available in the open Web that might complement the data that you are given.

Data Accessibility

Summarize how the data can be accessed. For example, data may be available for download in files, or accessible through an API, or can be queried from a database. Mention any restrictions in accessing the data, for example if it is sensitive data that can only accessed with special permission.

Data Formats

Describe briefly the formats of the data. Common data formats include CSV, JSON, XML, shapefiles, or any other specific formats relevant to your website.

Data Challenges

Summarize why analyzing this data will be challenging. This may include issues like data heterogeneity, data size, and any pre- or post-processing needs. Explain some ideas for how these challenges could be addressed.

Data Visualizations and Highlights

Including a visualization is a simple way to show something interesting about the data. Perhaps the visualizations could simply highlight the size, distribution, and other simple statistical characteristics of the data.

Problem and Requirements

Introduction

This page serves as a **Problem and Requirements Document** for the project. It should be no more than 5 pages long. It should be done after or in combination with the Data Assessment document. It should have an initial release after no more than four weeks into the project, and can serve as an interim project report. It can be refined as the project progresses and the project is better understood.

Motivation

The project advisor should have provided the general context for the project, the kinds of problems that could be addressed using the data available, and the overall motivation for working in this area.

Problem for the Semester

While the motivation section typically describes the long-term reasons to tackle this project, the problem section should focus on the specific problem and goals selected for the current semester. It should be a realistic goal that can be accomplished with the resources available, and this typically means the data, the people, the time available. It is useful to be specific about the initial questions that will be addressed first, because those typically lead to a better understanding of what other questions could be tackled. This will help establish a clear scope for the project that will give everyone reasonable confidence of what the team can achieve. It is also very useful to describe what the team is not going to tackle and has agreed to leave for future work.

State of the Art

Provide a brief summary or survey of the possible approaches, advances, or tools that are currently available that could potentially be used to address the problem.

Design and Approach

Discuss the initial approach that the team will follow. This may include the key idea that the team believes could potentially work, and a high-level description of 1) the system to be built and a diagram of its components, 2) the inputs to the system and the task that it will do with that data, and 3) the outputs to be generated. Discuss also a baseline system that is simple and can be quickly built to address the problem in a reasonable way even if it has poor performance, so that you can clearly show improvements. Discuss also how those improvements will be demonstrated through metrics or other means.

Use Case Scenario

Show with a use case scenario with examples of who the ultimate users could be, what the system that you plan to build for the semester will do in that scenario. Provide a mockup of the outputs for the use case that you propose.

Desired Outcomes and Benefits

Connect back the system that you will be building with the initial motivation for the work, and what additional future work might be needed in order for the system to provide benefits to its users.

Approach

This page contains key sections of the **Final Report** for the project focused on the data science methodology used to approach the problem. It should be no more than 3 pages long. It should be done after or in combination with the Requirements document. It should have an initial release after no more than eight weeks into the project, and can serve as an interim project report. It can be refined as the project progresses and the problem is better understood.

Data Quality

Describe any steps that were used to address any issues concerning the quality of the data. This may include collecting data quality metrics, discarding subsets of the data, or applying specific techniques for handling missing values, dealing with outliers, etc.

Data Preprocessing

Describe the steps taken to preprocess the raw data to prepare it for analysis. This may include data transformations to convert to a required format, feature engineering operations, encoding features as binary, etc.

Exploratory Data Analysis (EDA)

Discuss any techniques employed to gain insights into the data. This could include data visualizations, generating summary statistics, initial analysis, and other exploratory techniques used to understand the data distributions, features, and helpful patterns.

Model Development

Describe the algorithms, methodology, and architectures used to generate models. Discuss how models were generated, seeded, and improved. Show the libraries and frameworks used for model development, as well as the rationale behind those choices.

Model Evaluation

Discuss the evaluation metrics used to assess model performance, and justify those choices based on the problem that the project is addressing. Describe the evaluation techniques used, such as cross-validation, and how undesirable model behaviors, such as overfitting, were avoided.

Results

This page contains key sections of the **Final Report** for the project focused on results to date. It should be no more than 2 pages long. An initial draft can be created at any point during the project, and can be refined as the project progresses.

System and Model Performance

Show the performance of the best system and model(s) developed, showing clearly the performance metrics and improvements over the baseline system as appropriate. Create visualizations that show clearly these results.

Discussion of Findings

Offer a discussion of the main findings using the system developed. Put the results in the context of the original problem statement and the questions that were posed.

Discuss any unexpected results, and potential explanations.

Enumerate (ideally in bullets) the most important findings, and their impact on your project goals.

Limitations and Future Work

Discuss any limitations of the work to date, how these limitations could be addressed in future work. Discuss what lines of work are most promising given the understanding of the problem and the data gained throughout the project.