SOP: PSDS4900 Data Science Capstone

Synopsis

The Data Science and Analytics Capstone is a research or mission-driven project performed over an extended course engagement to demonstrate realistic execution of a full data science lifecycle, from raw data to mission insights, knowledge discovery, or automation. This project should include realistic data carpentry and exploratory analysis with more complex and/or variable data sources. The project will include significant, domain appropriate statistical or machine learning modeling for knowledge discovery, predictive analytics, prescriptive analytics, or analytical augmentation (automation).

Students should complete an applied data science research project that builds upon the knowledge and skills developed throughout the PSDS program. Students are paired with a faculty mentor and an NGA (or DoD/IC) stakeholder for the project. This is a 3-credit independent capstone project in Data Science. The capstone is an opportunity for the student to refine and polish their skills as a data scientist. The final capstone output should appropriately demonstrate the student's work and capabilities as a data scientist. The faculty member and NGA/DoD/IC stakeholder will serve as the project *mentors*.

Notes Regarding MU Credit

To receive MU credit and to satisfy certificate or degree requirements, a faculty member who serves as the graduate advisor for the student supervises the capstone. The PSDS team has various faculty members with appropriate clearances to serve on the examination committee.

Capstone Learning Objectives & Outcomes

- Students will learn to establish a data science project plan in concert with stakeholders. This should include identifying data, identifying, and refining the data question, and target exploratory outcomes.
- Students will perform a technical analysis on domain specific data, collaborating with domain expert (mentors) as needed to understand the contextual relevance of the data within a relevant problem space.
- Students will develop processes for the acquisition, manipulation, cleaning, and structuring of the data such that future and repeatable data integration is possible.
- Students will consult with the stakeholder to identify and potentially refine an open question of the data.
- Students will achieve proficiency in contextually appropriate data exploration, including both statistically and visually.
- Students will integrate disparate datasets into a comprehensive, complementary corpus that facilitates in-depth analysis and understanding
- Students will evaluate the suitability of the data for predictive analytics through

exploratory preliminary statistical modeling or machine learning.

- Students will evaluate a variety of predictive analytical methods to develop relevant decision-making products, such as mission-relevant intelligence or intelligent automations.
- Students will develop data stories that illustrate the journey from raw data to the decision- making products of their data science project.

Action Plan

Students should identify and select relevant data that may be in a variety of formats and conditions, and must work through the data ingestion, reshaping, and structuring process (data carpentry). Students will then integrate core and auxiliary datasets that enhance the contextual, qualitative, and quantitative exploration of a problem domain. Students will engage in exploratory data analysis, such as statistical investigation, data exploration through visualization, and preliminary exploratory statistical and machine-learning modeling. Students will develop a comprehensive plan to answer key stakeholder questions, as well as preliminary steps for data quality assessment and suitability for predictive modeling. To facilitate this as a collaborative effort, students with develop an action plan for the project, the plan will then be reviewed with the mentors for feedback and revision.

The plan should provide a logical path to the desired outcome products (see above). The plan should function as the framework for the project tasks and milestones, as well as a structure for project tracking and notes (progress, issues, resolutions, discoveries, etc.). These notes will undoubtedly help the student build a compelling data story as part of their final project deliverable.

Recommend Milestones:

- **Milestone 1:**
  Establish the project action plan. Identifying relevant data, formulate initial questions, and establish regular communications (telecon or zoom meetings) between student and mentors. Conduct initial/key data acquisition, begin data reshaping, cleaning, and structuring (data carpentry). The project's security classification will be determined here.

- **Milestone 2:**
  Establish Desired Outcomes. There should be specific questions you are seeking to answer with relevant data sets that you intend to work with. You should know enough about the data sets (i.e., *look* at them this milestone) to know what questions you are likely to be able to answer. This milestone should continue the data carpentry, setting the project up to move into exploratory data analysis.

- **Milestone 3:**
  Conduct Exploratory Data Analysis. This should include descriptive statistics

and basic visualizations. Additionally, the student should begin exploratory statistical modeling and machine learning.

  - ○ **<u>Deliverable 1</u>**: Initial descriptive statistics and/or visualizations; as well as summaries of exploratory statistical modeling/machine learning with assessments of suitability for data.

- **Milestone 4:**
  Identify additional datasets to enhance the likelihood of success. Detail how each additional dataset augments the existing data repository and will contribute to the contextual, qualitative, and quantitative understanding of the problem domain. Additionally, basic exploratory modelling to assess the utility of the auxiliary datasets.

- **Milestone 5:**
  Evolution of visualizations and computational models. Using feedback from your mentors and reflection on your progress, discoveries, and challenges thus far; refine and add specificity to your action plan. Begin identifying the structure and content of your project's Data Story.

- **Milestone 6:**
  Conduct extensive training and verification of predictive and/or prescriptive models. Based on the discoveries during the exploratory data analysis and modeling; students will train, validate, and test predictive and prescriptive analytical models that are relevant to the desired knowledge discoveries for the domain and the stakeholder. These models may include statistical models, machine learning, or data mining, just to name a few. The resulting models should be incorporated into the Data Story along with all the necessary caveats, assumptions, and uncertainties that are inherent based on the veracity of the data.
  - ○ **<u>Deliverable 2</u>**: A rigorous mathematical analysis of the statistical properties of the input data, the models (type, hyper parameterization, etc.), and the models' operating characteristics.

- **Milestone 7:**
  Develop your story with data. Recall the storytelling with data discussed in the *PSDS 1000 Introduction to Data Science* class and key concepts that were further emphasized in the *PSDS 2300 Data Visualization* class.
  - ○ **<u>Deliverable 3</u>**: A first draft of your story. You can communicate this via PowerPoint, PDF, or a website. The most important characteristic is that you include the description of: a) the questions or goals you have, b) why they are important, c) the suitability (perfections and imperfections) of the data you have chosen for achieving these goals, d) visual, data-driven drafts of how you will tell the story. If you do not know how to technically build exactly the visualization you have in mind for your project at this stage, then go ahead and sketch something out and share it.

- **Milestone 8:**
  Finalize your Data Story
  - **<u>Deliverable 4</u>**: Final Capstone Project Deliverable. Additionally, all your project intermediate work items (action plan, software, etc.).

Repeatability Requirement: All project work should be tracked using appropriate version control systems, such as GIT. Software developed to ingest, structure, model, and render the work products should be version controlled and reusable by the NGA community.

Capstone Calendar: Meetings, Milestones, Deliverables for OY4

Capstone Event	**Date**
Milestone 1: Capstone Kick-off	NCE – 18 Feb NCW – 19 Feb
Milestone 2 Email	March 12
Milestone 3 Email	April 2
Milestone 3 Feedback Meeting	NCE – 7 Apr NCW – 8 Apr
Milestone 4 Email	April 30
Milestone 5 Email	May 21
Milestone 6 Email	June 11
Milestone 7 Email	July 1
Milestone 7: Feedback Meeting	NCE – 7 July NCW – 8 July
Milestone 8 Email	Aug 5
Final Presentations	Aug 9, 2021
Final Presentations	Aug 10, 2021
Final Presentations	Aug 11, 2021

Important Note on Schedule: It is critical that students remain engaged in the Capstone process by communicating with their mentors and adhering to the milestone schedule. The schedule is designed to facilitate thorough and high-value projects of importance to the NGA enterprise or larger DoD/IC. Students that miss two milestones will be referred to the NGA College and dropped from the course. Prior arrangements for work-related TDY must be made ahead of time.