Capstone Project Charter

Project Name	Data Science Application in Marine Biology
Document Date	2/6/2015
Sponsor Name	Bill Howe
Sponsor Email	billhowe@cs.washington.edu
Sponsor Organization	eScience Institute, University of Washington

The Project Charter is a living document, and any substantial changes to the scope, work plan, deliverables, or stakeholders should be documented below, and an updated charter re-submitted to Canvas.

Revision History		
Date	**Author**	**Description of change**
06-02-2015	Tanmay Modak	Wrote the first third of the document after collaboration with team and sponsor.
06-02-2015	Elton Dias	Wrote the second third of the document after collaboration with team and sponsor.
06-02-2015	Abhigyan Kaustubh	Wrote the final third of the document after collaboration with team and sponsor.

The Project Overview should very clearly set boundaries for the Capstone Project. Early conversations with the project sponsor should cover all of the below elements of the project, clearly setting expectations between the sponsor and the project team.

Project Overview	
Project background	Oceanographers use flow cytometry to measure the optical properties of a given sample of water through radial dispersion. This is done by attaching flow cytometers to the bottom of ships that conduct research, thus enabling coverage of a vast body of water. We plan to collect flow cytometry data obtained at 3 minute time intervals (might change), and use a suitable clustering technique to identify regions in the water body that have similar trends in microscopic life form population.
Objectives	1. To map the different regions of a water body based on the variation in the population of different life forms in these regions.

	2. To identify regions within the water body that are similar/different to other regions based on the forward scatter data obtained using the SeaFlow instrument. 3. To draw meaningful inferences about the variation of the population of these life forms (and possibly correlate these inferences with environmental factors like temperature, salinity, etc.) 4. To successfully implement our algorithm in relevant environments (initially in Python, and potentially Myria and/or other large-scale systems)
Impacts & critical success factors	Success factors: 1. Sign-off from the science stakeholders on our model of the problem and intended solution 2. The insights that can be obtained by observing the data and implementing these clustering algorithms. 3. A preliminary clustering solution that can identify specific regions of water in a small, contrived sample of the data. 4. A full-scale clustering solution using Myria or other platform that identifies patterns across multiple cruises. 5. Scientific insight derived from this exercise, as defined by the science stakeholder. 6. Delivery of all code, examples, and documentation required for researchers to extend our methods, observations and inferences.
In scope	1. Identifying a clustering technique for the given data problem and using the insights to determine the marine ecosystem for a given area of sampling. Inferences on the ecosystem will be determined by the organisms only with no utilization of external factors like salinity, temperature, currents…etc.
Out of scope	1. Utilizing additional dataset features like salinity, temperature,etc to draw conclusions over the marine ecosystem as well as understanding migration patterns of the marine organism.
Stakeholders	1. UW Oceanography Researchers (eScience Institute) 2. Bill Howe (eScience Institute, UW CSE)

	3. Sophie Clayton, UW Oceanography
Key deliverables	1. A Python script that can run PCA on an individual SeaFlow 3-minute file 2. An initial solution that clusters the results of multiple SeaFlow files, and the results of an experimental evaluation to recover "known" clusters in a contrived subset of data provided by the oceanographers 3. A full-scale implementation of the method, and the results of processing a multi-cruise dataset 4. All code, examples, and documentation delivered in a github repository for future use. 5. Visualizations of the data and the results.
Schedule of deliverables	02/19- Identifying principal components using a small part of the dataset. 02/26: A high level analysis of Myria for the purpose of our project. Other deliverables are subject to the above two deliverables.

The project resources outlined below should be considered in concert with the expectations for the Project Overview above. Consider the number of team members and the approximate hours each member will have to allocate to the project, as well as access needed to other resources, such as hardware, software, or people (e.g., your sponsor).

Project Resources		
Project team & email addresses	Abhigyan Kaustubh	akhuia@uw.edu
	Elton Dias	eltond@uw.edu
	Tanmay Modak	tmodak@uw.edu

Capstone Project Charter

Hours and cost estimate	Hours : 10 hours/ week for this quarter 20 hrs/ week for the next quarter Cost estimate: 0	
Other resources: software, hardware, other equipment, or workspace	Myria, Python, R, Numerical Python Packages (Numpy, Scipy, Pandas)	
Sponsor time commitment (weekly)	5 to 10 hours (includes weekly update meetings, advice and email correspondence.)	
Sponsor role	Adviser/Overseer	

Project team members may have little control over other factors that affect project work. Think carefully about assumptions the team has made (e.g., “the sponsor’s database will be available to us offsite”), the constraints placed on the team (“the sponsor is not always able to reply to email over the weekend, when we are working”), and dependencies the project has on other events (“our sponsor uses exclusively Microsoft products, and therefore our solution must be compatible in X software package” or “the product must be in a form that enables transmission live to remote sites or users”).

Factors Affecting Project Work	
Assumptions	1. The flow cytometry data would be easily discernible after some basic reading on characteristic features of Marine Biology. Moreover, the data can be easily structured using a mixture of packages from commonly used Data Processing software. 2. The data science problem we are trying to address can be approached using standard machine learning techniques, or ensemble variations of commonly known techniques. 3. The three minute window should provide enough data to identify and label the marine population.
Constraints	1. Algorithm selection/development would be a top constraint considering the nature of individual 3min data snapshot as well as overall matching of similar characteristics over multiple snapshots. Hence identifying the marine ecosystem over an entire voyage. 2. Scalability complications of the developed solution with regards to deployment over large Big Data Systems for parallel processing.
Dependencies	1. Quality of Dataset to Analyze. 2. Availability of Computing Infrastructure to enable Analysis on scale. 3. Knowledgeable resources (Domain Experts) to acknowledge the correct application of computing constructs to classify the data.

Risks may also be under the team’s control, or not. In the below grid, indicate risks to carrying out any of the project’s expectations, as outlined above. Rank each risk in terms of Low, Medium or High (L, M, H) Probability and Impact. This will enable the team to proactively deal with risks as the project is carried out.

Risk Management Plan			
Risk	**Probability (L, M, H)**	**Impact (L, M, H)**	**Mitigation steps**
This project doesn't work in Myria as expected	M	H	Research other systems and be up to date with the pros and cons of each system and ways to approach the problem using these systems (Spark, Haloop, etc)
We might not be able to customize the algorithms effectively as needed	M	H	Research alternate clustering techniques, and be up to date with the pros and cons of each of these algorithms
Since this project involves new science, we cannot predict the outcome. If we cannot identify relevant biogeochemical patterns we will not know whether the method was insufficient, the patterns do not exist, or the data did not capture them.	L	H	NA

Capstone Project Charter

In order to ensure client satisfaction and fulfillment, it is crucial to communicate effectively with the project sponsor. This way, problems can be addressed early by the group. Each of the team members, and the sponsor, should understand the communication frequency (e.g., weekly on a certain day, after deliverables are complete, etc.) and agree upon its terms outlined below.

Communication Plan					
Team meeting frequency	Bi-weekly				
	Email	**Phone call**	**In-person meeting**	**Formal Report**	**Other (specify)**
Sponsor-preferred update format	Most preferred		Yes	Yes	
	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Reporting frequency (indicate day of week and how often – weekly, biweekly, etc.)				In person meeting with the sponsor	
Alternate point of contact at sponsor organization (email)	Jeremy Hyrkas <hyrkas@cs.washington.edu>				
Issues management plan	1. For smoother team operation: a. Regular in person meetings to touch base with progress achieved and future plan of action. b. Constant individual research of machine learning techniques as well as the problem domain (marine biology) c. Produce key deliverables on a weekly/bi-weekly basis to endure consistent progress. 2. If issues arise: a. Maintain flexibility in terms of project scope. b. Research alternate methods to arrive at a viable solution.				

	c. Approach sponsor for intervention. d. Research alternate technologies and plans of action for the same as a backup.
Change management plan	1. To prioritise certain tasks over others keeping the broader picture in mind. 2. To try different algorithms/approaches as we get a better understanding of the dataset. 3. To try to deal with different scales of implementation (terabyte/gigabyte) 4. To broaden or narrow the scope (we might factor in environmental variables, for example, or discard some of the questions we are trying to answer in the interest of realistically trying to address our core issue)

Project Co-ordinators: Abhigyan Kaustubh and Elton Dias