Requirements Specification

December 13th, 2019

Canopy - Team 11

Team Members:
Robert Plueger
Nicholas Lopez
Dongyu Xia
Maria Granroth
Sponsor: Dr. Patrick Jantz
Mentor: Scooter Nowak

This document has been accepted as the baseline requirements for the project.

Team Lead: ______________________ Date: _______

Sponsor: _________________________ Date: _______

Version 1.2

Table of Contents

1 Introduction

We are Canopy, a 2019 - 2020 Northern Arizona University Computer Science capstone team. Our team members are Robert Plueger (team lead), Maria Granroth, Nicholas Lopez, and Dongyu Xia. Our project is to develop an application for characterizing 3D vegetation structure in tropical ecosystems. We will analyze the data obtained from the Global Ecosystem Dynamics Investigation (GEDI). GEDI data is collected via lidar, laser ranging of forests and topography. The data consists of the 3D distribution of stems, branches, and leaves, making up the vegetation structure of ecosystems. GEDI is an Earth Ventures mission that started in December 2018 and is funded by NASA. GEDI will acquire billions of vegetation profiles across the Earth's temperate and tropical ecosystems in two years.

The sponsor and client of this project is Dr. Patrick Jantz. He is a member of the Vegetation Structure as an Essential Biodiversity Variable (VSEBV) project based at NAU. He is also a team member funded by the NASA Applied Sciences program. He will use GEDI data to develop vegetation structure essential biodiversity variables (EBVs). EBVs can be used by policy makers and scientists to improve land use decisions and guide priorities for conservation of biodiversity in tropical landscapes.

The reason for launching this project is that the current workflow is clumsy and requires a lot of manual work. The goal is to create an application to help improve processing speed and accuracy while reducing repetition and manual steps. We envision an application that supports both researchers and policy makers. High quality analysis from our web application will make it easier for our end users to make decisions for conservation and resource management.

2 Problem Statement

Our client's current workflow is manual and coded in R. Our client would like GEDI data analysis to be accessible for everyone. With its current shortcomings, most users would not be able or willing to analyze GEDI data. To better understand our client's current workflow, please refer to *Diagram 1*. Please note that every step of the workflow requires constant user interaction.

This system is failing to meet expectations as follows:

- The user must directly process the data with R
    - Users may not be able to use R
- Analyzing the data is currently tedious and complex
    - The user must locate relevant NASA files by hand
    - Files must be loaded and processed one at a time, requiring constant user attention
- The output is provided in an esoteric layout
    - Some users may find the analysis difficult to understand

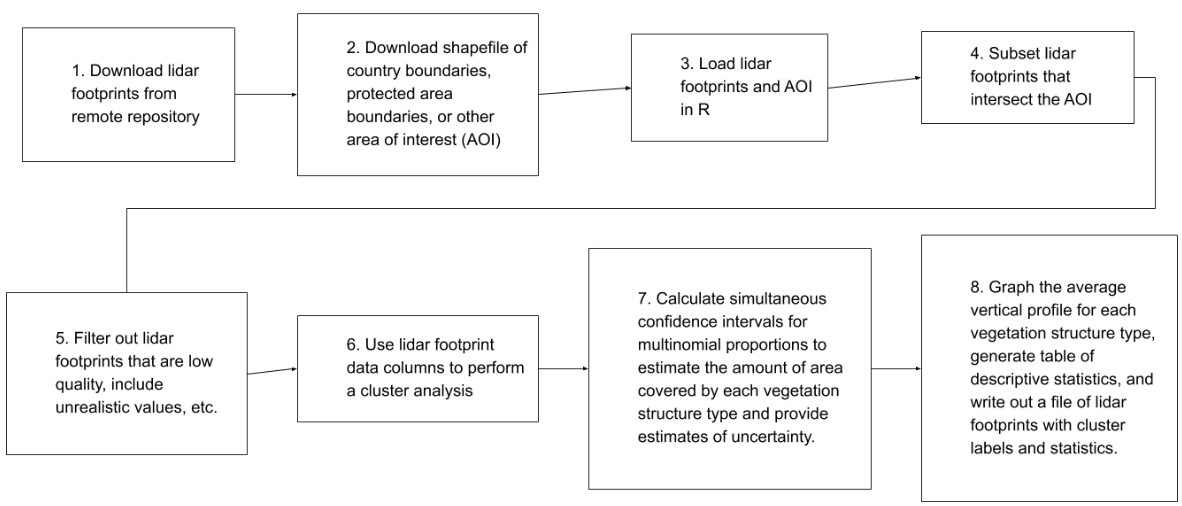

Diagram 1 – Current Workflow

3 Solution Vision

Our solution is to build a web application that allows users to understand the forest structure for a region that they choose. It will analyze the data for them based on the statistical functions they choose and help them understand the results. We will be doing the following as key project features:

- Convert existing code to Python scripts that run autonomously
    - The user will provide the area they wish to analyze and the statistical functions they want performed
- Create a system for automatically requesting GEDI data from NASA repositories
    - Users will select individual regions to be analyzed while our system handles the shape files
- Process this data into a useful state
    - The system will analyze the data for the user and group it into logical partitions
- Present the analyzed data to the user in an easy to understand way
    - A data guide will be provided alongside the completed analyses

Reference *Diagram 2* to see a workflow based on our solution. This workflow is what the user will experience. The user only has to know what they want, and the system will do it for them. Compared to the current workflow shown in *Diagram 1*, this is much easier for the user.

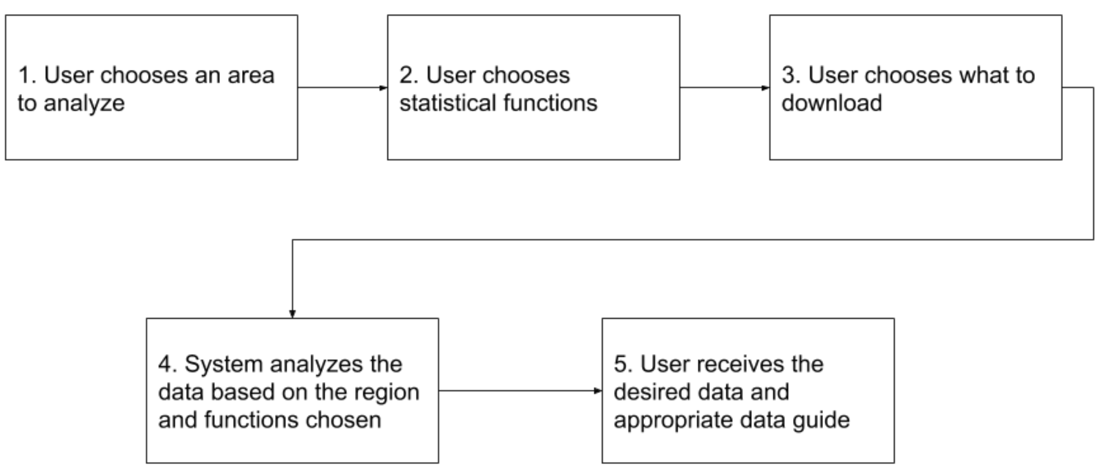

Diagram 2 – Solution Workflow

4 Project Requirements

This section will detail the complete requirements for the system described in our *Solution Vision*. These requirements will guide the process of building the web application. Requirements were acquired through client meetings performed roughly every other week.

The purpose of this web application is to make statistical analysis of GEDI data easily accessible to a wide audience. In order to achieve this, domain-level requirements need to be established. These requirements can be described as what the user needs from the web application in order for it to fulfill its purpose. The domain-level requirements for this project are as follows:

1. Our solution must download and analyze GEDI data
2. Our solution must allow users to specify an area to analyze
3. Our solution must allow users to choose a variety of statistical functions
4. Our solution must provide the analyzed data back to user

These requirements will be broken down further in the following sections.

4.1 Functional Requirements

The domain-level requirements introduced in the last section can be expanded into functional requirements. These are the functions that the team must provide in order for the project to be considered complete. Please refer to *Diagram 3* for a hierarchical representation of how these requirements are broken down into more detailed requirements. Refer to the use case on the same page to understand how all of the requirements will come together to form a coherent system. This section will describe each of the requirements shown.

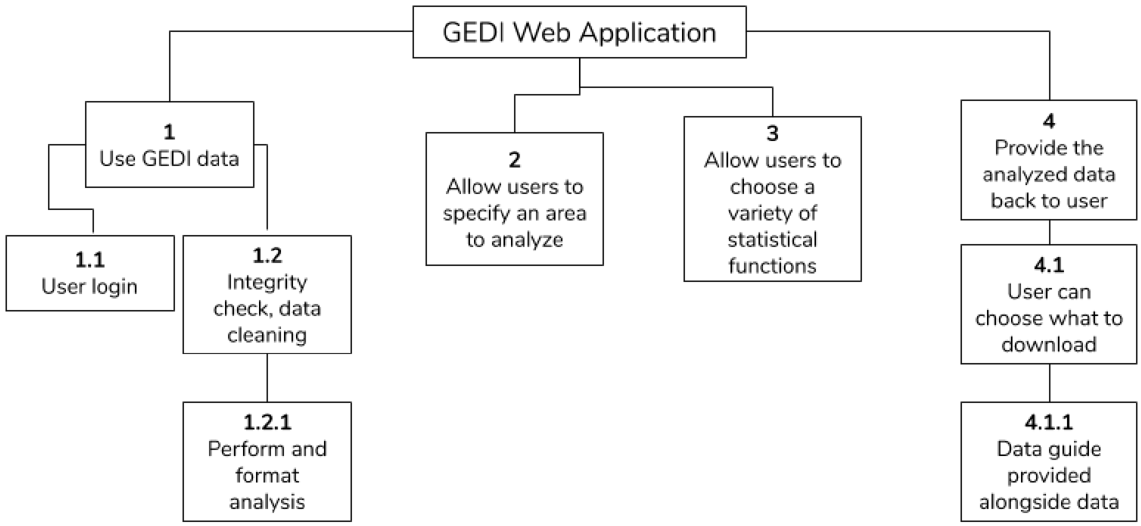

Diagram 3 - Hierarchical Decomposition of the Requirements

Primary expected use case

Requirements: 1 - 4
Precondition: User is on the website
Postconditions: The data is downloaded from AppEARS, analyzed, and returned to the user.

1. User logs in **(1.1)**
2. User specifies the area to analyze **(2)**
3. User specifies statistical function(s) desired **(3)**
4. User chooses what to download. **(4.1)**
5. User confirms that the things chosen are correct.
6. The system indicates that it has received the information.
7. The system cleans any data received from the user. **(1.2)**
8. The system downloads the GEDI data pertaining to the chosen area from AppEARS. **(1)**
9. The system cleans the GEDI data. **(1.2)**
10. The system performs the chosen functions. **(1.2.1 & 3)**
11. The system packages the things chosen to be downloaded. **(4.1)**
12. The system chooses the relevant data guide(s) to the forest structure type. **(4.1.1)**
13. The system returns the data package and its guide to the user. **(4)**

4.1.1 Our solution must download and analyze GEDI data

The web application must use the most up-to-date GEDI data available to the public.

4.1.1.1 Our solution will implement a user login

In order to download and analyze GEDI data, NASA requires an Earthdata user login. There are two ways that the project can approach this. Either the web application can use Earthdata's Login API to have the user log in, or if that fails, it can use the project client's login to grab the GEDI data on the user's behalf.

4.1.1.2 Our solution must perform an Integrity check and data cleaning

The system should check the GEDI data flags to see if the data is appropriate to use. If the user provided a shapefile, then the system should clean it if it is not formatted correctly. This is to ensure that the analysis is performed as correctly as possible.

4.1.1.2.1 Our solution must perform and format the analysis

Once the GEDI data is received and the user has specified the area and statistical functions they wish to use, the system must perform and format the analysis appropriately.

4.1.2 Our solution must allow users to specify an area to analyze

Users must be able to specify the area that they wish to analyze. This should be done so that the user has plenty of options available to them, which allows flexibility for what they need. For example, this could include a map, menu, and an option where the user can upload their own shapefile that specifies the area.

4.1.3 Our solution must allow users to choose a variety of statistical functions

Users must be able to choose how they want the area analyzed. They may choose one or several functions to have performed on the area that they specified. These functions will provide basic summaries of the data found in the area specified. Examples of the functions that a user may choose from are listed below. This list may not be exhaustive, but the functions currently listed will be included.

1. **Number of Observations -** The total number of lidar pings that were performed in that area

2. **Data Quality -** The average quality of the observations
3. **Density of Observations -** The amount of observations per km2
4. **Standard deviation of the density of observations** - Standard deviation performed on the density of observations
5. **Mean Vegetation Height** - The average vegetation height found over the entire area
6. **Standard Deviation of Vegetation Height** - The standard deviation of the vegetation heights found over the entire area

4.1.4 Our solution must provide the analyzed data back to user

Once the system has analyzed the data, it must provide that data back to the user for download. The user may use the resulting data however they want.

4.1.4.1 Our solution must allow the user to choose what to download

The user must be able to specify what they want downloaded. This could include the original GEDI data or the analysis by itself.

4.1.4.1.1 Our solution must provide a data guide alongside data

There must be a graphical guide provided to the user to help explain the analysis results. This will include a representation of the different forest structure types. In addition, a representation of the corresponding forest structure variables (e.g. height, cover, vertical distribution) for each structure type will be provided. The guide will also provide an estimate on how accurate the structure type is based on the best guess of how many different structure types are found in that area.

This guide will be produced by the system via data driven clustering based on the variables mentioned.

We are confident that we will be able to meet these functional requirements by May 2020. Performance requirements, where applicable, will be discussed in the next section.

4.2 Performance Requirements

Not every web application does everything at the same speed. Some processes take longer than others naturally and vary from platform to platform. Our project has requirements about how well and how quickly some components of our system should perform.

We have two main components of our system to take into consideration regarding speed. Both the statistical analyses and user downloads will require different processing time. However, as shown in *Diagram 4*, both of the processes start with the same steps, represented by the green arrows. The process of sending a request to our web application to then request data from NASA should be able to complete in a matter of seconds.

4.2.1 Requirements of Statistical Analyses

(Functional Requirements 1.2.1, 2, 3, & 4)

The statistical analyses steps are represented by the blue arrows in *Diagram 4*. The process of receiving data from NASA, analyzing the data, and sending it to the user takes a variable amount of time. This amount of time depends on how many files are received from NASA. The analyses should take roughly one hour if 50,000 files are received. As each request will be different, we will only be able to estimate that an analysis may take up to three days if the information is requested for an entire country / continent.

4.2.2 Requirements of Data Downloading

(Functional Requirements 4.1 & 4.1.1)

The user's data downloading process is represented by the orange arrow in *Diagram 4.* The amount of time this process takes is more difficult to control, as it is dependent on the user's internet speed. It is difficult to determine the exact time that it will take to download data. Different download speeds will require different amounts of time to complete the data transfer.

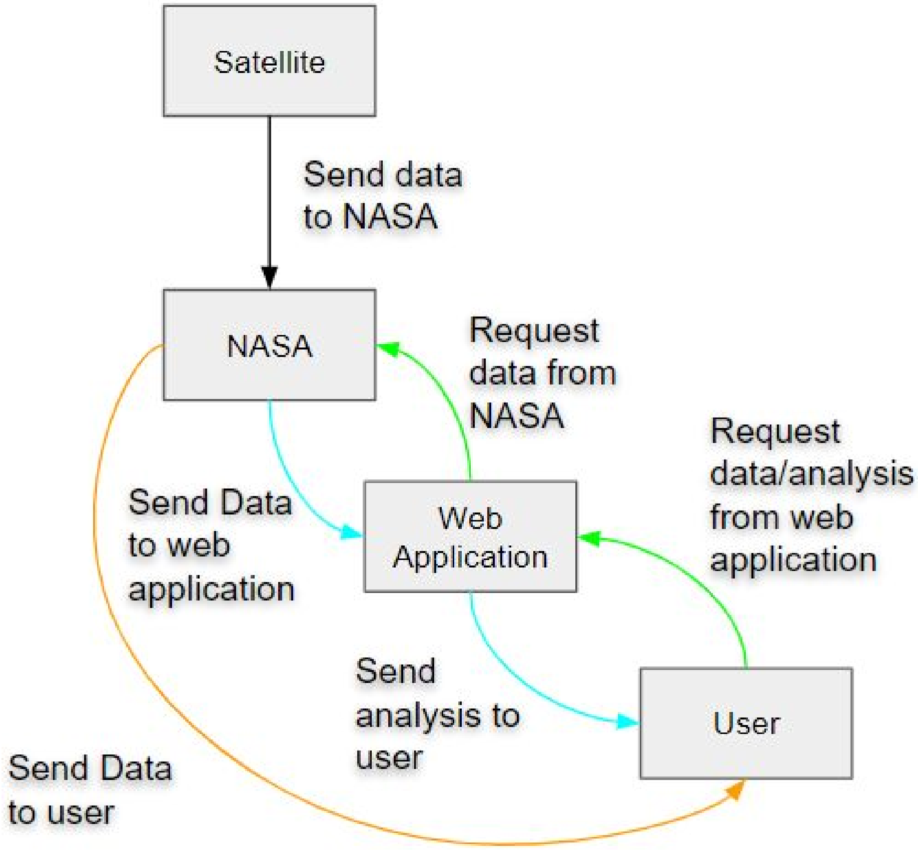

Diagram 4 - Process of User Requesting Information

4.3 Environmental Constraints

There are some things that can limit or tighten the bounds of a project. These limitations can be imposed by the client, hardware, or software compatibility. Our project has two areas where there are constraints: primary programming language and web hosting service tier.

4.3.1 Constraints on Primary Programming Language

Our client has expressed his desire to use either Python or R as our programming language. This is due to his familiarity with these languages and his desire to continue this project after our collaboration.

4.3.2 Constraints on Web Hosting Service Tier

(Functional Requirements 4, 4.1, & 4.1.1)

We are additionally constrained by the project's budget. A portion of the budget will be devoted to hosting our web application. Hosting a web application through Microsoft Azure requires a regular payment. The price paid determines the amount of storage and connection speed (tier) our application can have.

5 Potential Risks

There are some challenges we may encounter that could create risks for our project and our end users. These risks differ based on which user comes across these challenges.

5.1 Risks to Researchers

Researchers may use our website and its data to gather information for a report. Such a report could be published in scientific journals, magazines, or as a standalone research paper. A challenge that we can come across is producing incorrect analyses of data. If our analyses produce incorrect information, then a researcher risks using this information to publish incorrect findings. This can result in the diminished credibility of that researcher and our website. Additionally, another researcher could use that published information and publish something else. This would create a chain of misinformation and lost credibility. Although this is moderately severe, the likelihood of this happening is quite low, as we can do extensive testing to make sure our analyses will be correct. *Diagram 5* gives a visual summary of the risks for researchers.

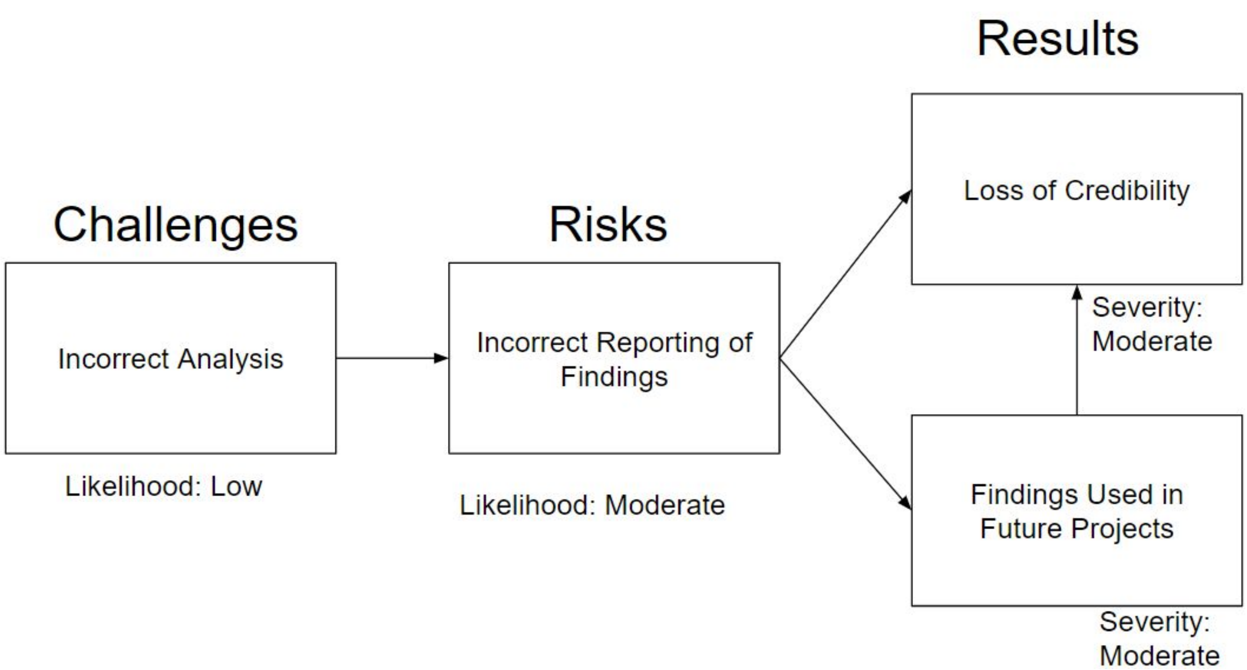

Diagram 5 - Researcher Challenges, Risks, and Results

5.2 Risks to Policy Makers

Policy makers may use our website and its data to determine the best course of action for certain environmental areas. They may decide to increase conservation efforts of an ecosystem, costing money and resources. They may also decide to let an ecosystem be destroyed or removed if it is decided to be less valuable or less sustainable.

We have the same challenge as before, potentially producing incorrect analytic information, but we also have the challenge of providing inadequate information. This stems from only knowing about the vegetation of an ecosystem but not the residing wildlife. Both of these challenges create the risk of a policy maker deciding the wrong environmental action to take. They may decide to destroy an ecosystem that is very valuable or they may decide to increase conservation efforts on an ecosystem that is not valuable or sustainable. The likelihood of providing inadequate data is moderate, as we cannot prevent policy makers from taking action, aside from warning them that wildlife may be present. The severity of destroying a valuable ecosystem is severe: there could be global environmental impacts. Protecting a less valuable ecosystem has a low severity, as there will be no environmental damage, but resources will be used without profit. *Diagram 6* gives a visual summary of the risks for policy makers.

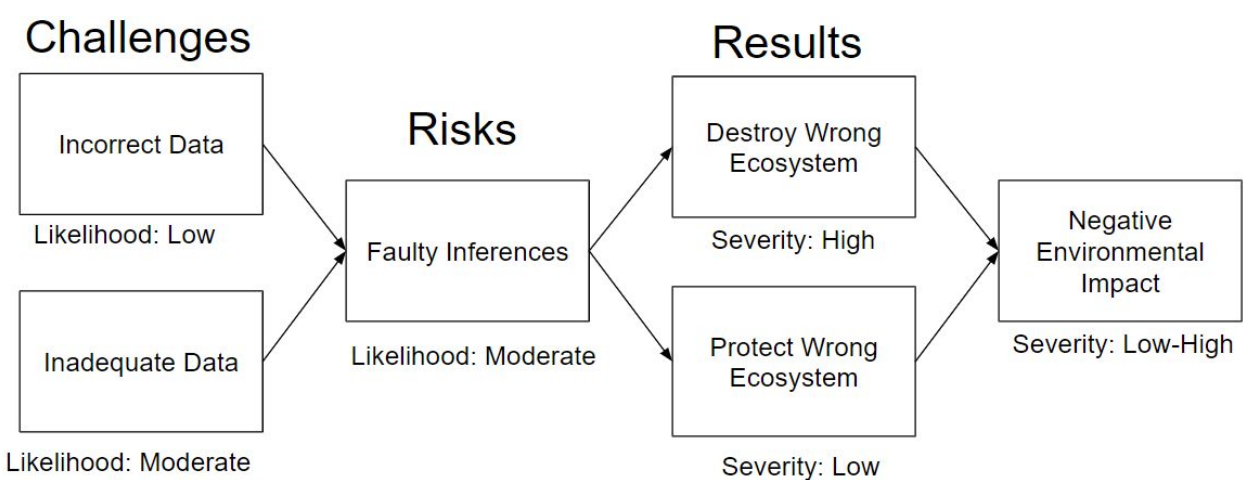

Diagram 6 - Policy Maker Challenges, Risks, and Results

6 Project Plan

This section will briefly discuss our project execution plans as it stands right now. A Gantt chart will describe some milestones in our plans for the next phase.

6.1 Execution Plan

After many exchanges with our client, we have determined the client's needs. We are currently completing the Requirements Specification document based on information received from the client. At the same time, we are also preparing a Technical Prototypes Demonstration. We are making this demo according to the requirements of the client. This demo will have a basic GUI, can perform basic mathematical operations on virtual data, and will upload and download files through the GUI.

6.2 Next Phase

For next semester, we designed a rough schedule, pictured in *Diagram 7*. In February, we plan to finish the function that will allow the user to specify an area and the data analysis they want. Then, in March we are going to finish the data guide and download functions. In April, we will finish the login function and communicate in depth with Dr. Jantz about the product. In May, we will improve the product, and try to develop a perfect application.

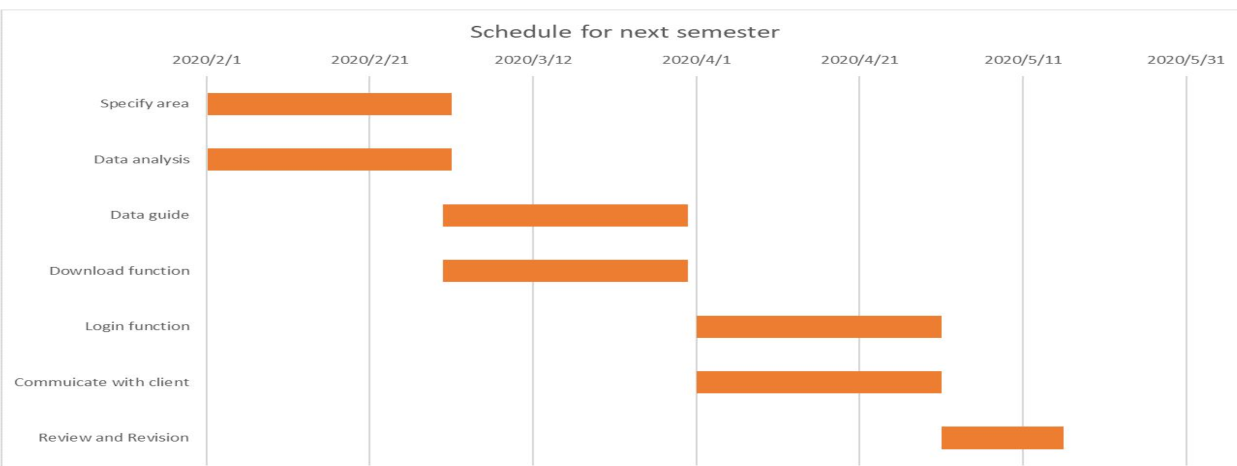

Diagram 7-Schedule for Next Phase

7 Conclusion

Our client, Dr. Patrick Jantz, is a NASA Applied Sciences team member that will use GEDI data to develop vegetation structure essential biodiversity variables. These variables can be used by researchers and policy makers to improve land use and conserve the biodiversity of tropical landscapes.

We discovered that our client's current process is manual and complex, which is not ideal for general users. To improve this, we will create a web application that allows the user to specify the area and the data analysis they want, while the system does the analysis for them and provides the resulting data in an easily accessible form. Our web application will feature a system that will grab the GEDI data for the user based on the region specified and Python scripts to analyze that data. Our application will group the data into a useful state and present it back to the user with an accompanying data guide to help the user understand the results.

Our project is currently on track. We are confident that the plans discussed in this document and the schedule designed for next semester (January through May 2020) will keep us on track to develop a useful product that satisfies Dr. Jantz's criteria.