**Department of Computer Science**

**City University of Hong Kong**

**CS4514 Final Year Project**

**Project Plan**

**Title: Big Data Analysis on Ecological Data**

**Project Code: 17CS095**

**Student Name: Lam Wing Ting**

**Student ID: 54038173**

**Supervisor: Dr. Wong Ka Chun**

**Table of Contents**

[1. Introduction 3](#_Toc494103408)

[1.1 Background Information 3](#_Toc494103409)

[1.2 Aims and Objectives 3](#_Toc494103410)

[1.3 Scope 3](#_Toc494103411)

[2. Methodology 4](#_Toc494103412)

[2.1 Data Collection 4](#_Toc494103413)

[2.2 Data Analysis 4](#_Toc494103414)

[3. Major Technical Components 5](#_Toc494103415)

[3.1 Data Mining 5](#_Toc494103416)

[3.2 Web Application 5](#_Toc494103417)

[4. Expected Results and Deliverables 6](#_Toc494103418)

[5. Project Schedule 7](#_Toc494103419)

[6. Reference 8](#_Toc494103420)

# Introduction

### Background Information

Refer to Biology online dictionary, ecology is a science concerned with the interactions of living organisms with each other and with their environment. The statistics published by Census of Marine Life in 2011, estimated that there are around 3 million to 100 million species ranging from the land to ocean depths on earth. Besides, there are still unknown living organisms not seen by eyes. Investigation of ecosystem has been conducted by researchers for decades. Researchers spent their life long time to study any external, internal or even unknown factors affecting the ecosystem.

As a result, sinking in thousands of sparse and large data sets, it is not easy to conduct the research and obtain the results in a short period of time. In fact, pattern of researching on species can be very similar. Since the species belong to the same classification, they have similar features and also their attributes in the data sets. For example, several studies focus on birds but the attributes of data are quite similar or even the same e.g. temperature, humidity, types of bird, weather, etc.

### Aims and Objectives

The project aims to identify the similar and common data analysis practice among ecological data sets. By performing data mining on different data types like compilation, experimental, observational and time series, similar and common data mining models might be identified.

To visualize the data analysis result, a web application will be developed with API supports i.e. deck.gl developed by Uber and Google Map API. Using the APIs, the application will compare data among different data sets and display on the map. User is able to select the timespan of the data for comparison. Then, depending on the types of enquiries, display the result using maps or the charts.

When the data mining model is identified, the web application can provide some guidance to researchers. Thus, researchers can spare the time for further and deeper data analysis in their study. Also, deck.gl is newly developed by Uber since 2009 and made available in 2016. If the web application can integrate this API, it can improve the data visualization with large data sets and some interactive user events.

### Scope

Since there are more than thousands of data sets, not all data sets will be explored. Mainly focuses on the data sets of similar species with similar attributes. The data sets also include the location and coordinates for map visualization and route tracking.

# Methodology

### Data Collection

Data are mainly selected from Ecological Data Wiki and Data.Gov.HK for analysis.

Ecological Data Wiki is a web platform for scientists providing their data sources around the world. In the website, ecological datasets can be found quickly with brief description of each data set e.g. timespan of the data, label of the attributes, objectives of the study, etc. Users submit their data files or any quick links on the platform which visitors can easily access to the ecological data. Wide variety of species is available ranging from amphibians to microbes.

Another source is from Data.Gov.HK. It is public sector information portal of HKSAR. Datasets are provided by government departments and public/private organizations. The portal divides several data categories and in this project, will select the data sets under environment which is provided by Agriculture, Fisheries and Conservation Department.

### Data Analysis

Python is used for data mining with open source tool called IPython with Jupyter Notebook. Association and clustering analysis will be performed in every selected data sets.

For association analysis, frequent itemset generation will be used to identify high occurrence of data attributes. For the study about route tracking, high occurrence of combination of observation points can be identified and figure out the migration route of species.

For clustering analysis, K-Means clustering will be used. In the ecological studies, it is very common to set up observation points in different locations. With the location coordinates, calculate the centroids and indicate the place with high occurrence of species or other results.

Upon completion of individual data analysis, evaluate the flow of data analysis of each data set. Then, figure out a data analysis model which can cater for similar data sets. Frequent itemset generation is also used which can quickly identify the common attributes used by the researchers. It helps to figure out the methodology of common ecological study.

# Major Technical Components

### Data Mining

Weka

Weka is a open source software developed by The University of Waikato and in full name which is “Waikato Environment for Knowledge Analysis”. The project team has included several standard machine learning techniques into Weka. It can be used as data pre-processing, classification, clustering, association and visualization.

IPython with Jupyter Notebook

IPython is common to use in machine learning and provide a comprehensive environment for interactive computing. Jupyter Notebook is open source web application to share documents containing code and visualizations. It also handles data cleaning, data transformation and machine learning. It supports over 40 programming languages e.g. Python. This is the reason why using IPython with Jupyter Notebook.

### Web Application

Client Side – HTML, CSS and Javascript

The web application is implemented by HTML and Javascript and CSS for the style of web pages. Basic elements of websites can be achieved by HTML, CSS and Javascript.

Server Side – Python

Since Python is also used for data mining, server side will also be written in Python. It also helps the integration of Python library.

Database Server – MySQL

Data sets will be managed by MySQL since some data formats are designed for MySQL.

Data Visualization API - Google Map API and deck.gl by Uber

Google Map API is developed by Google and supports Javascript which is convenient for the web application. Deck.gl is developed by Uber since 2009 and designs for visualization of large data sets. It gives impressive visual results with limited effort and provides a complete architecture with Javascript support. These two APIs will be used for map visualization and display the data analysis result.

# Expected Results and Deliverables

A common work flow of data analysis used in ecological study will be identified . A web application is developed for data visualization and give guidance with the common work flow. The application also integrates with deck.gl or Google Map API to display on maps. Some interactive user events will be added , for example, users select the timespan of data and the application will generate some comparison results using charts or maps.

# C:\Users\tcpsa\AppData\Local\Microsoft\Windows\INetCache\Content.Word\FYP Project Schedule (6).pngProject Schedule

# Reference

1. Weka: <http://www.cs.waikato.ac.nz/ml/index.html>
2. Deck.GL by Uber: <https://uber.github.io/deck.gl/#/>
3. Ecological Data Wiki: <http://ecologicaldata.org/>
4. Data.Gov.HK: <https://data.gov.hk/en-data/dataset/hk-afcd-afcdlist-red-tide-location>