Project Report Guidelines

First and foremost, following the individual original work policy clearly stated at the start of the course, the topic and questions you ask in your project must be of your own invention. **If you used ideas from a particular web site or previous project, or did your project as part of an existing research collaboration, you must identify your sources and/or collaborators and provide links and citation(s) where appropriate.**

As a guide, the report should be around 8-10 pages depending on space used for any visualizations, tables, etc.

The format of the report is semi-flexible - you can include additional information, but at a minimum it should have the following sections:

# Motivation

1. Motivation (5 points):  (a) Briefly state the nature of your project and why you chose it. (b)  What specific question or goal did you try to address?

Our proposed project was to use Argo data (a data set of ocean floats that collect variables such as salinity and temperature in oceans around the world), map the change of salinity and temperature in the Estuary and Gulf of St. Lawrence, and identify potential correlations to fish populations totals in the same area.

The Gulf of St. Lawrence is an incredibly diverse and complex marine and estuary ecosystem and is one of the largest of its kind in the world. The area is made up of freshwater from the Canadian Shield, the Great Lakes basin and the St. Lawrence River system emptying out into the Atlantic Ocean, where it combines with the cold Labrador Current from the Arctic and the warm Gulf Stream from the tropics. All of these currents merge in a semi-enclosed and mostly shallow area, creating the perfect conditions for incredible diversity of life. The Gulf of St. Lawrence is the most important source of fish on the Atlantic side for the commercial fishing industry in both the U.S. and Canada, both top exporters of fish and seafood in the world[[1]](#footnote-1). By studying and reporting about this ecosystem and the species that depend on it we hope to better inform others about its critical importance.

* How has ocean properties such as temperature, salinity, etc., changed over a period of ten years in the Estuary and Gulf of St. Lawrence?
* How have fish populations (pelagic primarily, but also possibly demersal fish populations as well) changed in the Estuary and Gulf of St. Lawrence within the same time period?
* Is there a meaningful correlation between temperatures/salinity/other factors and population that merits further study?

# Data Sources

2. Data Sources (10 points):  Describe the properties of the two dataset(s) or API services you used. Be specific. Your information at a minimum should include but not be limited to:

* where the datasets or API resources are located,
* what formats they returned/used,
* what were the important variables contained in them,
* how many records you used or retrieved (if using an API), and
* what time periods they covered (if there is a time element)

For example, if you downloaded data or used API services, you should state the specific URLs to those files or resources. It should require zero effort on my part to find and access the exact resources you used if I need to do so.

## Argo Data

### **Name**: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) –Atlantic Ocean, 2009-2019

### **Short description**: Argo is an international program that collects information about Earth’s oceans using a fleet of robotic instruments that drift with the ocean currents and move between the surface and mid-water level. These floats capture information such as salinity, temperature, oxygen levels and depth. For the scope of this project, we will use the Atlantic Ocean dataset from 2009 to 2019.

### **Size**: approx. 20 GB

### **Location**: <ftp://usgodae.org/pub/outgoing/argo>

### **Format**: netCDF files

### **Access method**: FTP and from relational database post processing.

## Global Temperature and Salinity Profile Programme Data

### **Name:** Global Temperature and Salinity Profile Programme (GTSPP) data

### **Data Location**: FTP - ftp://ftp.nodc.noaa.gov/pub/data.nodc/gtspp/best\_nc/

### **Formats used**: individual netCDF files aggregated in a monthly tarball

### **Important variables contained in each file**: latitude, longitude, station or ship identifier, timestamp of measurement, depth of measurement (in meters), salinity, and temperature. Supplementing all these measurements are additional measurements denoting the quality of the measurement (e.g., ‘good’, ‘probably good’, ‘probably bad’, etc.).

### **Time Periods Covered**: 2009-2019 (although the dataset goes further back).

### **Size**: 132 tar.gz files, 7.83 GB (compressed)

Based on an initial exploration of the ARGO dataset, it became quickly clear that the Argo dataset might be relatively sparse on float data within the Estuary and Gulf of St. Lawrence. Data from the Global Temperature and Salinity Profile Programme (GTSPP), a joint international cooperative effort supporting the World Climate Research Prorgramme, was used to supplement ARGO data for the same time period (2009-2019). Similar to Argo data, GTSPP data also collects sea measurements such as salinity and temperature in the Earth’s oceans, but instead of floats, data is collected from both ships and buoys.

## Department of Fisheries and Oceans (DFO) – Quebec Coastal Thermograph Network

### Name: Department of Fisheries and Oceans (DFO) - Quebec Coastal Thermograph Network

### Data Location: <https://open.canada.ca/data/en/dataset/848e943b-1a98-43b8-acb3-ac89af17ea41> (HTTP access)

### Formats Used: a zipped file containing 2 folders, 1 with CSV files of each station and another folder with graphs.

### Important variables contained in each file: Timestamp of measurement, station id, latitude, longitude, depth of measurement, temperature and salinity.

### Time Periods Covered: The data set includes 1980-2019, but we used a filtered subset of this dataset.

DFO’s dataset was used to supplement both the Argo float and GTSPP dataset at the surface level, as we weren’t sure we would have sufficient measurement points for the Estuary and Gulf of St. Lawrence from the first two datasets. Data are collected from buoys and, unlike both Argo and GTSPP data, strictly surface level data only (< than 100 meters).

## Pelagic Fish Populations in the Estuary and Gulf of St. Lawrence

### **Name**: Pelagic fish species abundance in the Estuary and Gulf of St. Lawrence between 2009 and 2018

### **Data Location**: <https://open.canada.ca/data/en/dataset/f1fc359c-0ed1-4045-a421-adef2497b68d>

### **Formats Used**: a zipped file containing 2 CSV files from http download, one for each location.

### **Important variables contained in each file**: Date of measurement, the station id, latitude, longitude, and quantity of specific fish species.

### **Time Periods Covered**: 2009 - 2018

### **Size**: approx. 1,700 lines of data per file (0.2MB zipped)

### **Access method**: HTTP download and from relational database post processing

The Canadian Department of Fisheries and Ocean conducts annual multidisciplinary surveys of the Northern and Southern Gulf of St. Lawrence to capture information on groundfish and invertebrates’ abundance, spatial distribution and diversity. The pelagic species represented in the dataset are: Arctic Cod, Atlantic Argentine, Atlantic Herring, Atlantic Mackerel, Atlantic Soft Pout, Capelin, Lumpfish, Pollock, Rainbow Smelt, Sand Lances, Silver Hake, Threespine Stickleback and White Barracudina.

## Estuary and Gulf of St. Lawrence Shape File

### **Name**: Gulf of St. Lawrence shapefile

### **Short Description**: Shapefile of the Gulf and Estuary of St. Lawrence for visualization purposes.

### **Size**: <1MB

### **Location**: https://www.marineregions.org/gazetteer.php?p=details&id=4290

### **Format**: Shapefile

### **Access method**: Download and open and manipulate with Python’s shapely and/or geopandas.

## Ocean and Fisheries Data

# Data Manipulation

3. Data Manipulation Methods (30 points):  For each of your two sources, describe how you manipulated the data.  For example:

* How specifically did you need to manipulate the data?
* How did you handle missing, incomplete, or incorrect data?
* How did you perform conversion or processing steps?
* What variables and steps did you use to join the two data resources to perform your data analysis?
* Briefly describe the workflow of your source code and what the main parts do.
* What challenges did you encounter and how did you solve them?

## Data helpers/packages (files are located in the /helper folder)

Since our datasets are quite large and not conducive towards local Jupyter manipulation at large (neither of us, sadly, have unlimited memory), our data processing and manipulation jobs were chunked and we used a variety of helpers and processes to move data for cleanup and visualization.

### db.py: Since our data is stored in AWS RDS, a helper was written to move data to and from the databases.

## Argo

## GTSPP Data

GTSPP data is quite large (7.8GB while still zipped!), so a number of challenges were presented. Almost all were solved via writing an extract-transform-load process

### How to access the data? Data were stored in individual tarballs via FTP.

## Department of Fisheries and Oceans (DFO) – Quebec Coastal Thermograph Network

## Pelagic Fish Populations in the Estuary and Gulf of St. Lawrence

## Estuary and Gulf of St. Lawrence Shape File

# Analysis and Visualization

4. Analysis and Visualization (25 points):

* A key goal of this project was bringing together two different data resources to answer an interesting question or find a new insight that could not have been answered with either data resource alone (which you summarized in part 1).  Now describe the analysis steps you performed on your combined dataset to address that goal/question. Be specific, and include references to key functions or parts of your code.
* What interesting relationships or insights did you get from your analysis?
* What didn't work, and why?
* To summarize your findings, include at least one visualization (chart, plot, tag cloud, map or other graphic) that summarizes your analysis.

# What Worked

# What Didn’t Work

# Statement of Work

5. Statement of Work (0 points)

* You must include a statement that describes the contribution that each team member made to the project.

Sharon Sung

1. <https://www.dfo-mpo.gc.ca/stats/trade-commerce/world-mondial/export/wxv1517-eng.htm> [↑](#footnote-ref-1)