Draft Proposal Guidelines

You should use the following outline for your proposal, which shouldn't need more than about one page.  The proposal draft (which will be reviewed by your peers; see below) is worth up to 10 points; the revised proposal (which will be reviewed by the course instructors; also see below) is also worth up to 10 points.

1. Summarize your proposed project in a few sentences.

* What is your proposed project and why are you proposing it?
* What are the question(s) you want to answer, or goal to achieve?

# Project Summary

## Summary

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

## Questions

### How has ocean temperatures and salinity changed over a period of ten years in the Estuary and Gulf of St. Lawrence?

### How has the fish populations (pelagic primarily, but also possibly demersal fish populations as well) changed in the Estuary and Gulf of St. Lawrence in the same time period?

### Is there a meaningful correlation between temperatures/salinity/other factors and population to further study?

2. Describe two different data sources you plan to access, manipulate and bring together.  The data sources must require different access mechanisms and/or use different data formats. (For example, you might pick one data source that uses a Web API that returns JSON, and the second might use SQL to query a database, or fetch and parse an HTML page.)

For each data set, you should summarize these properties:

* Name
* Short description (i.e., 1-3 sentences)
* Size (in records and/or bytes)
* Location (give the URL or other access method)
* Format
* Access method

# Data Sets

## Data Set 1

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

### **Short description**: Argo is an international program that collects information from inside the ocean using a fleet of robotic instruments that drift with the ocean currents and move up and down between the surface and a mid-water level. It captures information such as salinity, temperature, oxygen levels and depth. For the scope of this project, we will use the Atlantic Ocean data from 2009 to 2019.

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

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

### **Format**: netCDF files

### **Access method**: FTP and import to relational database post pre-processing.

## Data Set 2

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

### **Short Description**: The Canadian Department of Fisheries and Ocean has been conducting 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, Three spine Stickleback and White Barracudina.

### **Size**: approx. 1,800 lines of data (0.2MB)

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

### **Format**: CSV /REST API

### **Access method**: Download and import to relational database via pymysql

3. Describe with 1-3 sentences for each point below what data manipulation is likely to be needed:

* What initial processing will have to be done on each?
* How will you combine the datasets, and what will be produced as output?
* What new information will result from combining them?

# Data Manipulation Needed

## Initial processing:

### The Argo Atlantic Ocean data will need to be queried and filtered to retrieve data from close to the Estuary and Gulf of St. Lawrence. NetCDF files will need to be transformed into CSV files to be stored in a table in a relational database. We plan on using Amazon RDS for our project.

### The Pelagic fish data will need to be downloaded and uploaded as a table in our relational database.

## Data manipulation:

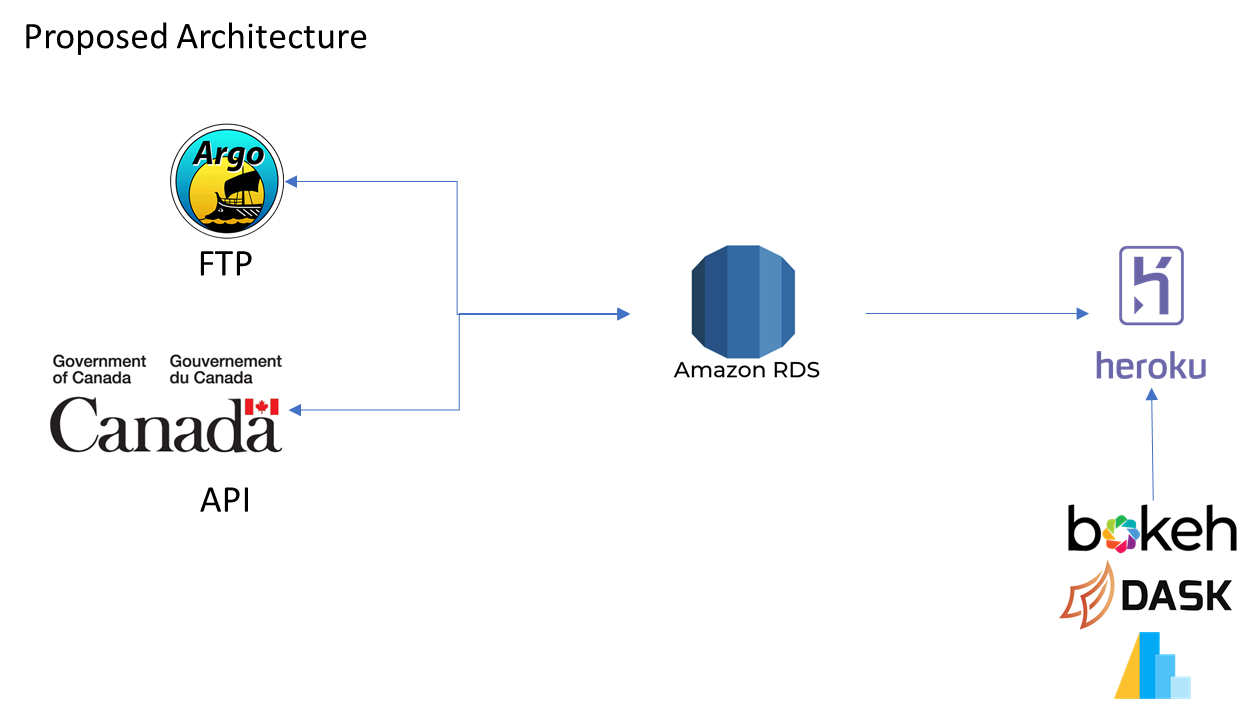
### Fishes and floats are not necessarily in the same exact location and depth, we will need to define shapefiles and depth bins on which we can join our datasets.

### Since the pelagic fish data is annual, we cannot compare it with our monthly ocean data. We will need to compute annual average temperature, salinity, oxygen level and their changes compared to the previous year. We can also compute the year-to-year increase/decrease of the pelagic fish data.

### Then we can join our data sets on shapefiles and depth bins.

## Output:

### With the joined dataset, we can create several maps to compare variations in ocean data at different depths with fish populations. To this end, we plan on using a combination of Dask and Bokeh (or Altair) to build an interactive dashboard to visualize the impact of ocean data variation on each fish population over time (i.e. do increases/decreases in temperature, salinity and oxygen level impact pelagic fish population, and if so, do they impact them in the same way?). This dashboard would be made publicly available via the Heroku platform.



4. Describe in 1-3 sentences one interesting visualization you could include in a final presentation and report that would show the value/answer a particular question in your final output dataset (that would not be possible with either of the original datasets alone).

Maps are a must when dealing with geographic data and should be included in our final presentation and report. An interesting visualization would be to map overall changes in temperature and salinity over the 10-year period our dataset covers (2009-2019) with overall changes in fish population to visualize potential correlations and identify potential thresholds that can endanger specific fishes. For example, a one-year variation of +0.01°C may not have a significant impact, but an overall increase of +0.5°C over 10 years may drastically impact the Arctic Cod population.

5. Indicate the contribution that each team member will make to the project.

Sharon Sung:

Claire-Isabelle Carlier: