JupyterBook Migration and Machine Learning with Sea Floor Sampling Video

Mentors: Matt Biddle(@MathewBiddle), and
Dalton Kell(@daltonkell) , Matt Iannucci, Tara Franey, Stephanie Berkman, Joe Zottoli

---

Personal Information

Name: Munakala Lohith

University: Indian Institute of Information Technology Kalyani
Course: Bachelors of Technology in Computer Science and Engineering
Course Term: 4 Years (2018-2022)
Current Year: 3rd Year

Email: lohithmunakala@gmail.com
Phone: +91 630 440 5087 (India)

GitHub: lohithmunakala
LinkedIn: Lohith Munakala
Medium: Lohith Munalala
Resume: Link

Time Zone: IST(GMT +5:30)

Synopsis

1. **Machine Learning with Sea Floor Sampling Video**

   In this project, the main goal is to reduce the manual work that goes into visually processing the vessel mounted seafloor videos which are hundreds of hours long and are expensive and time consuming to do. The reduction is done using deep learning models for object detection and image classification. Tracking is also a focus of this project so that unique species can be identified. There is no existing repo for the project but the idea is almost similar to VIAME but also focuses on the environment in the background of the images (eg. sand, gravel etc) ie. the identification of the habitat. The focus is also to calculate the total area that is surveyed along a particular transect. The results that are expected from this project are to be saved in a standardised format so that they can be easily integrated into many of the reports.

2. **JupyterBook migration for IOOS Data Demo Center**

   The main aim of the project is to migrate all the tutorial IOOS notebooks from the old website to a newer JupyterBook Framework. *JupyterBook is an easy to use open-source project which helps build publication-quality books and documentation from computational material.* All the notebooks lie in the IOOS Notebook Demo repo. The migration should encourage newer authors to publish more notebooks into different chapters as this would be beneficial to anyone using the IOOS servers and data to get an idea of how the data is collected and processed. The process should be automatic such that any new notebook included in specific chapters should be processed and the new jupyterbook website should be generated. The whole process of completing this has been mentioned in the Deliverables section below.

Benefits to Community

1. The **Machine Learning SeaFloor Sampling project** would help IOOS and the open-source community to use the object detection and tracking models to reduce

the time taken to process seafloor videos, enhancing the erddap servers with fresh and instant data. The data could also be used for many other purposes like helping the environment which is one of the priorities of IOOS. The project will make the repetitive process more intuitive and will make the overall seafloor habitat visualization more efficient.

2. The **IOOS Data Demo Center migration** to JupyterBook would make the discovery of the demo notebooks easier and simplified as the notebooks will be classified into various chapters. It will also make the addition of newer notebooks easier to understand and implement for various researchers/contributors to add or create new demo notebooks that will help the community.

Deliverables

Under the project here are the changes and improvements I intend to do in the 10 weeks of the program.

1. **Machine Learning with Seafloor Sampling Video**

I will be creating a new repository that will be used for all the additions and creation of models. The repository will have specific files such as .sh and .bat files which will automate the process of detection of each of the following methods. These methods are based on what is proposed here.

- ➔ Collection and annotation of data
  - ◆ The data will be initially collected and if the data is unannotated, the data will be later annotated. The annotation will be done using LabelImg with the help of biologists.
  - ◆ If the data is already existing, this specific step can be skipped.

- ➔ Image Enhancement
  - ◆ As the data will be noisy and unfiltered, I intend to first clean the data so that training can be easy and the same will be applied to the incoming data.
    - Contrast enhancement
      - The images will be made clearer using techniques such as OpenCV's adaptive histogram equalization. This will increase the dynamic range of the videos/images.
    - Illumination normalization
      - The goal of this process is to increase the gamma values of the image to a specific set of values. An example of this process is mentioned here.
    - Colour correction

        - Colour correction will be done using OpenCV's colour correction module.

➔ Identification of species:
- ◆ Implementation details are as follows:
    - Using YOLOv3 + darknet
    - Why darknet? Darknet is an efficient backbone for YOLOv3 because of the following specification
        - It requires only 5.58 billion operations
        - It achieves 72.9% top-1 accuracy and 91.2% accuracy on ImageNet
        - It has an efficient architectural combination:
            - ◆ 3*3 filters to extract features
            - ◆ 1*1 filters to reduce output channels
            - ◆ Global average pooling
    - The dataset that will be used here will be an already annotated dataset mentioned above.

➔ Identification of habitat
- ◆ Implementation details are as follows for the habitat detection:
    - A CNN will be trained to identify the various background ie. habitats of the various species like sand, gravel, shells etc.
    - The dataset that will be used will be decided by the mentors as this is a major part of the project.
    - Based on the dataset and training results, a combination of GoogLeNet and VGG16 the final prediction will be as they have impressive performance on ImageNet.

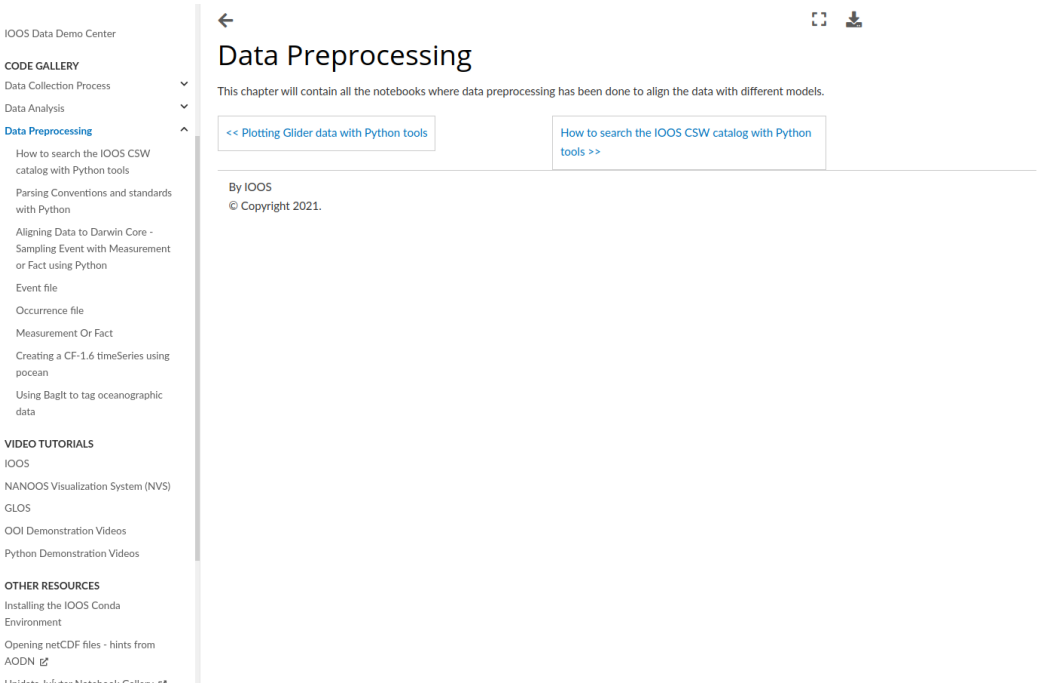
➔ Generation of CSV files from the above processes.
- ◆ The CSV/metadata from all the above processes will be merged into one single CSV file.
- ◆ The data from the object detection model, the habitat detection and the text recognition model will be combined into one single entity based on the timestamps,

2. JupyterBook migration for IOOS Data Demo Center

I have already submitted a PR where I have done most of the work of the project. All the work that has been done is based on what is mentioned here. The rest of the additional process that will be added is mentioned below here.

➔ Formatting the existing notebooks

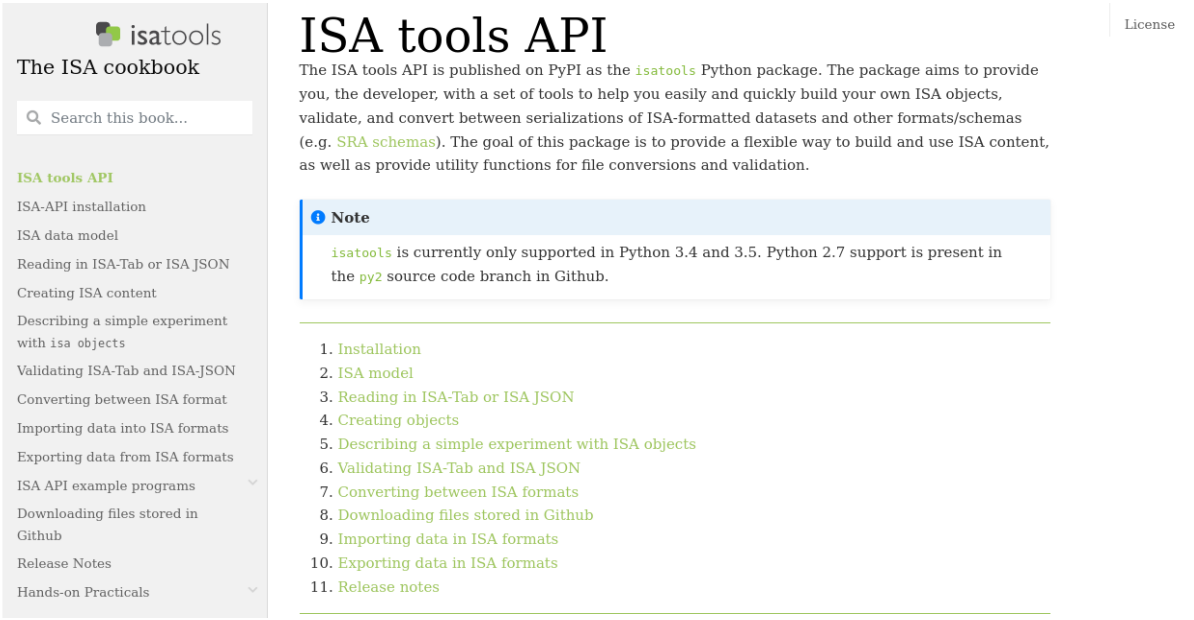
- Some of the notebooks have formatting issues. For example, the Aligning data to Dalvin Core has some Markdown issues where the reading sizes are different and this leads to the generation of additional unnecessary content for the jupyterbook chapters.



In the following example, we can see how the subheadings of the notebook have generated additional content in the chapters which is unnecessary. In order to automate it, we would need to define a set of rules which people contributing will follow.

➔ Beautification of the existing jupyterbook

- The existing jupyterbook in the PR can be beatified according to the standards requested by the IOOS team. The colours and the fonts, size of fonts could be changed, which can be done after the approval of the team.
- An example could be the following jupyterbook.



➔ Automating the process of addition of notebooks
- The main aim of this project, after the completion of the above-mentioned tasks, the next task would be to automate the pages such that every time a new notebook is added to the repository in its respective folder, the _toc file gets updated automatically and the website is updated with the new notebook.
- I plan to explore both github pages as well as netifly and implement the best out of the two. This would be the best option as I get to explore and implement the best of both worlds considering each other's pros and cons.

➔ Document the jupyternotebook process
- Documentation to be added to help contributors improve and make changes to the jupyterbook in the future.

Tentative Timeline

Here is a detailed timeline of the work that I intend to do. I have divided the columns for both the SeaFloor Sampling Project and the IOOS JupyterBook project and intend to do them parallelly. This is possible as I have a vacation during the initial part of the coding phase and intend to work for 20 hours a week.

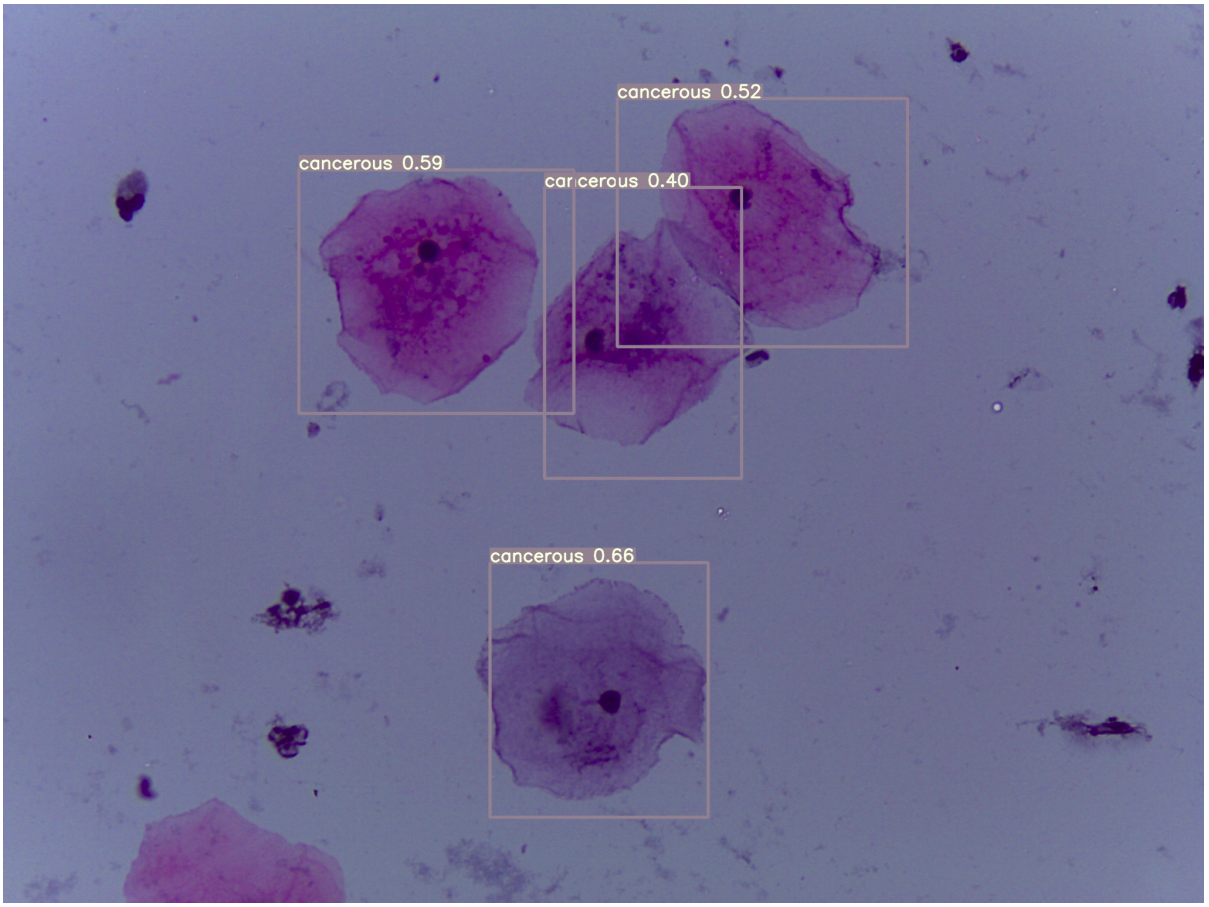
TIME	WORK ON SEAFLOOR SAMPLING VIDEO	WORK ON IOOS JUPYTERBOOKS
Before GSoC (April 13- May 17)	● Learning more about how seafloor videos are sampled and collected.	● Work on formatting the notebooks and solving the remaining issues.
Community Bonding Phase (May 17-June7)	● Get to know the mentors and the community well. ● Collecting the data from various sources. ● Looking at various methods of enhancing the seafloor data.	● Get to know the mentors and the community well as the two projects have a different set of mentors. ● Read the documentation of jupyterbook to enhance what has already been done.
Week 1 (June 7 - June 14)	● Annotating the data such that it can be used for object detection and image recognition. ● Collecting the data for text extraction and habitat classification.	● Enhance the jupyterbook with newer formats (in terms of colours used etc.)
Week 2 (June 14- June 21)	● Perform various enhancing methods to the video data such as debayering, colour correction, illumination normalization, and image contrast enhancement.	● Continue enhancing the jupyterbook.
Week 3 (June 21 - June 28)	Habitat Classification Model Training and Deployment ● Downloading and cleaning the habitat dataset ● Training 2 CNN models for habitat classification (VGG16 + InceptionV3)	● Find out ways to automate the process of adding new notebooks to the jupyterbook
Week 4 (June 28 - July 09)	Habitat Classification Model Deployment ● Ensembling the two models into one ● Deploy the model	● Experimenting with ways to automate the process of adding new jupyter notebooks
Week 5 (July 05 - July 12)	Object Detection Model Training ● Downloading and cleaning the ImageNet and whatever dataset we have Training a YOLOv3 with darknet as the backbone for species classification and detection.	● Ensuring all the implementation of the above-mentioned steps ie. ○ Formatting ○ Enhancement ○ Automation Take place properly without any issues before the first evaluation.

	Ensuring all the steps have been implemented before the phase 1 evaluation	
Week 6 (July 12 - July 16)	Phase 1 Evaluation	Phase 1 Evaluation
Week 6+7 (July 17- July 26)	Object Detection Model Deployment ● Deploy the object detection model for identifying different species in videos.	● Implementing automation for the jupyterbooks
Week 8 (July 26 - Aug 02)	Object Detection Model Deployment ● Deploy the object detection model for identifying different species in videos. Script for getting data from the model to a CSV file ● Write a script to get the final CSV(data) file so that all the extracted data is meaningful.	● Continue with the implementation for the jupyterbooks
Week 9 (Aug 02 - Aug 09)	● Continue with writing the script for the CSV and performing checks	● Check integration with the rest of the database ● Adding documentation to the project. ● Unit Testing
Week 10 (Aug 09 - Aug 16)	● Buffer week for documentation and completion of the remaining works if any ● Discussion with mentor	● Buffer week for any pending works ● Discussion with the mentor
	Final Evaluation	Final Evaluation

Related Work

1. Sea Floor Sampling Project:

- **Object Detection:** I have plenty of experience in object detection as I have been making a couple of models for object detection for the past 6 months for PAP Smear Analysis, which is a cancer detection project and I have done it using YOLOv5. I can use the basics from this project in the detection of aquatic species and marine life for images and videos. Here are some of the training I did on Colab and the results of object detection.



The results are pretty good too. This gives me a headstart as I know the basics of the training a simple YOLO model and could scale this up to videos of marine life.

- **Habitat Classification:** I had done a project based on classifying different groups of people based on the images. I used a simple CNN based transfer learning approach and the accuracy I got is 92%. The GitHub link is here. I used a custom Learning Rate finder and using a cyclic learning rate to train this model and this resulted in excellent results.
  - I believe that I could use the same experience of training groups of people into classifying the habitats from the images and videos that will be present.

- **NLP(Text extraction and Recognition):** I solved an intent-based text classification problem using BERT and used it to gather information about files content and classify and store the information in the database itself. The link to the GitHub repo can be found here. I believe this small part of NLP would give me a headstart for the extraction and the recognition part.

2. IOOS Jupyterbook

- I have already submitted a PR where I have created a jupyterbook for the IOOS demo centre with various chapters in it. The chapters are divided into

    - Data Access
    - Data Analysis and Visualization
    - Data Publication

  I have pushed all the notebooks into the chapters and have gone through the notebook_demo repository pretty well in the past several weeks. I have made **10 commits** to the PR and have kept changing the jupyterbook as per the requirements specified by Matt and Filipe.

Why I am the right person for this project:

Firstly, I believe I have the right set of skills for these projects as,

- I have taken the following courses in my Undergraduate
    1. Machine Learning
    2. Image Processing and Computer Vision
    3. Artificial Intelligence
    4. Programming in Python
- I have done projects which are based on
    1. CNN based Classification
    2. Object Detection using YOLO
    3. OpenCV and Docker

Secondly, I have been in constant touch with the mentors of both of the projects and my continuous commitment towards the projects can be seen there. I have discussed various issues and ways I have consistently solved the problems I have faced.

After GSoC

I want to continue contributing to the community even after the GSoC phase ends as these projects being new, will have a lot of improvements in terms of scale and features that can be added over time. As GSoC is only a 10 week period, additional improvements can be included after the GSoC period is over.

A few examples would be as follows:

- Creation of a GUI for the seafloor sampling project as this would be easier for the biologists to use the application.
- Containerizing the model so that it can become easier to deploy on a server so that it can be ready to use as and when necessary.
- Detection of text from the video (if any). This could be an addition to recognizing the texts using a CRNN and using SynthText++ respectively.

I genuinely believe that collaboration is the key to solving the upcoming problems and I believe this organization is solving some of the key problems in the world and being a part of this community would be my honour.

Conclusion

I would like to thank you for taking out the time to go through my proposal. Hoping to work with this amazing community.