|  |  |
| --- | --- |
| **MODULE 5: DATA CURATION AND VISUALIZATION OF EO DATA FOR MACHINE LEARNING** | |
| **OBJECTIVES** | * Understand general data exploration and cleaning methods * Understand exploration and processing techniques for general image data * Demonstrate theoretical and applied knowledge of image processing techniques to clean EO data * Demonstrate theoretical and applied knowledge of multi-spectral EO data visualization * Demonstrate applied knowledge of python for cross-platform EO data manipulation and visualization * Explain the key considerations for EO data manipulation with regard to machine learning |
| **METHODS** | Brainstorming, demos/code-alongs, application exercises, and structured discussion |
| **DURATION** | 12 hours for participants |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SESSION** | | | **DURATION** | **PARTICIPANTS…** |
| Presence-based | 5.1 | Introduction to data cleaning & processing | 1 hour | * Revise on the basics of data cleaning methodology and exploratory data analysis (EDA) methods in the ML context * Explain the data cleaning process for structured data * Explain numerical and visual methods for data quality checking |
| 5.2 | Data curation with image datasets | 1 hour | * Explain the different image pre-processing, augmentation, and EDA approaches * Explain the impact of image training data curation decisions on downstream ML processes |
| 5.3 | Data curation with EO data | 6 hours | * Explain the different types of approaches to correct EO images * Perform EDA on EO image data including summarizing an image dataset * Pull data from different EO platforms using different python libraries and utilities * Perform EO image data processing including cloud masking and removal |
| 5.4 | Visualization of EO data | 3.5 hours | * Describe the basic principles of EO multi-spectral imaging * Explain visualization methods for different analytical objectives * Apply band stacking in python * Visualize and explore EO data bands |

**OVERVIEW OF EXERCISES**

|  |  |  |
| --- | --- | --- |
| **#** | **Exercise Name** | **Alignment with Learning Objectives** |
| Ex 5.1 | **Land-Use-Land-Cover Pipeline using Sentinel Hub** | Use of Sentinal Hub via API (sentinelhub-py) with Rwandan geodata to:   a) Extract an AOI from a larger dataset   b) Apply cloud masking   c) Apply imputation (e.g. temporal interpolation) |
| Ex 5.2 | **Agricultural Pipeline using Google Earth Engine** | Use of Google Earth Engine (API) in python to visualize geodata:   a) Clip and filtering GEE image collections  b) Summarize and analyze time series   b) Apply reduction to image collections  c) Create mosaics |
| Ex 5.3 | **Cloud masking using Google Earth Engine** | Use of Google Earth Engine (API) in python to visualize geodata:   a) Identify cloud cover components and cloud shadows in EO data   b) process cloud cover   c) Remove cloud cover from map |
| Ex 5.4 | **EDA vizualization using EarthPy** | Use of earthpy package to visualize Rwandan geodata:   a) import S2 image data  b) build a band stack  c) vizualize S2 bands  d) vizualize indices built using S2 bands (NDVI, NDMI etc.) |
| Ex 5.5 | **Object detection using Google Earth Engine** | Use of Google Earth Engine to:   a) clip an AOI from an image collection and filter by time   b) apply canny edge detection  c) apply hough transform  d) compare to ground truth |
| Demo 1  CODE ALONG | **EDA on Landsat data using Earthpy** | Use of earthpy package to visualize geodata:   a) conduct band analysis of landsatl data   b) conduct RGB analysis   c) Intensity histogram correction |

**5.1** **INTRODUCTION TO DATA CLEANING & PROCESSING**

**Quiz questions**

|  |  |
| --- | --- |
| 1. **What are the common problems associated with data collection that can lead to data quality issues?**   (name at least 3) | * Missing values * Inaccurate measurements * Inconsistency * Duplicates * Bias |
| 1. **What are the general steps in data cleaning?**   (name at least 3) | - Screening  - Diagnosis  - Treatment  - Documentation |
| 1. **What do you use to create low-dimensional visualization of high dimensional data?** | * T-SNE |
| 1. **What are some EDA techniques for multiple features?** | * Scatter plots * Heatmaps * PCA * Correlation |
| 1. **What are some EDA techniques for image preprocessing?** | * Image resizing * Reorienting * Color correction |
| 1. **Mention some techniques for image data EDA.** | * Image inspection * Image statistics * Image summarization * Image class distribution * Dimensionality reduction * Color analysis |

**Videos**

* Exploratory Data Analysis (5:01): <https://www.youtube.com/watch?v=QiqZliDXCCg>

**Handouts**

Refer to separate Powerpoint presentation: Introduction to data cleaning and processing

**5.2** **DATA CURATION WITH IMAGE DATASETS**

**Videos or Quiz questions**

* Image Processing – Erosion (0:49) <https://www.youtube.com/watch?v=fmyE7DiaIYQ>
* Image Processing – Dilation (1:01) <https://www.youtube.com/watch?v=xO3ED27rMHs>

**Handouts**

Refer to separate PowerPoint presentation: Data curation with image datasets

**5.3** **DATA CURATION WITH EO DATA**

**Application exercises in Jupyter Notebooks on data curation with EO data**

* **Exercise 5.1 Land-Use-Land-Cover Pipeline using Sentinel Hub** (Individual work)
  + In this exercise, you will learn the preliminary steps for data accessing and exploration to construct Machine Learning pipelines.  
     First, download the notebook from Moodle and upload it to your Google Drive. Then, open the notebook. Read all the instructions, comments and notes within the notebook thoroughly and work through the codes. There is one question that asks you to complete the code for visualization.   
      
    Question 5.1.1: Complete the code for visualization. Once you have displayed the figure on the notebook. Download the figure (Right Click > Save image as). Rename the downloaded image to 5\_1\_visualize.png
  + When you finish the exercise download your notebook from Google Drive (File > Download > Download .ipynb) for submission.
* **Exercise 5.2 Agricultural Pipeline using Google Earth Engine** (Individual Work)
  + In this exercise, you will learn how to configure GEE to work with data from different satellite platforms.   
    First, download the notebook from Moodle and upload it to your Google Drive. Then, open the notebook.   
    Read all the instructions, comments and notes within the notebook thoroughly and work through the codes. There are five questions that ask you different things and three discussion questions. The discussion questions do not require you to submit anything. You will use these questions to discuss solutions with your fellow learners. Here are the submission requirements for the 5 questions.  
     **Question 5.2.1**: Complete the code for the filtering task. Copy the code you wrote in the cell into a text file named 5\_2\_1\_code.txt for submission. Please submit only the code relevant to this specific question.  
     **Question** **5.2.2**: Complete the code for the EE image to Pandas Dataframe conversion task. Copy the code you wrote into a text file named 5\_2\_2\_conversion.txt for submission. The code should be a single function.  
     **Question 5.2.3:** Submit the dataframe you created as part of this question in CSV format. The CSV file should be named toa\_rw.csv .  
     **Question 5.2.4:** Complete the code to visualize the trend. Submit the code in a file named 5\_2\_4\_code.txt. Download the plot you generated, rename it to 5\_2\_4\_figure.png  
     **Question 5.2.5:** Complete the code to visualize the map and submit a screen shot of the map. Save the screenshot to a file named 5\_2\_5\_map.png.
  + When you finish the exercise download your notebook from Google Drive (File > Download > Download .ipynb) for submission.
* **Exercise 5.3 Cloud masking using Google Earth Engine** (Individual work)
  + In this exercise, you will learn cloud masking.   
    First, download the notebook from Moodle and upload it to your Google Drive. Then, open the notebook. Read all the instructions, comments and notes within the notebook thoroughly and work through the codes. There are 5 questions and one discussion question. The discussion question does not require a submission.  
     **Question 5.3.1**: Complete the code for the mask generation. Copy the code you wrote in the cell into a text file named 5\_3\_1\_code.txt for submission. Please submit only the code relevant to this specific question.  
     **Question** **5.3.2**: Experiment with 5 different sets of the parameters as specified in the question submit each screenshot with the naming format specified in the question itself.   
    Question 5.3.3: Complete the code to do the filtering task and submit it in a file named 5\_3\_3\_code.txt  
     **Question 5.3.4:** Complete the code to do the cloud masking and submit it in a file named 5\_3\_4\_code.txt  
     **Question 5.3.5:** Complete the code to visualize the map and submit a screen shot of the map. Save the screenshot to a file named 5\_3\_5\_map.png.
  + When you finish the exercise download your notebook from Google Drive (File > Download > Download .ipynb) for submission.

* **Exercise 5.5 Object Detection using GEE** (Group work)
  + In this exercise, you will learn how to use GEE and OpenCV to detect fields in the Vaalharts Irrigation.   
    First, download the notebook from Moodle and upload it to your Google Drive. Then, open the notebook. Read all the instructions, comments and notes within the notebook thoroughly and work through the codes.  
     There are two questions and five discussion questions. The discussion question does not require a submission.

**Question 5.5.1**: Experiment with the thresholds and download the image for each experiment for submission.  
 **Question** **5.5.2**: Complete the code to compute the NDVI computation, Canny Edge detection, and map drawing. Submit the code in a file named 5\_4\_2\_code.txt and the screenshot of the map in a file named 5\_4\_2\_map.png.

* When you finish the exercise download your notebook from Google Drive (File > Download > Download .ipynb) for submission.

**Handouts**

Refer to separate Powerpoint presentation: Data curation with EO data

**5.4** **Visualization of EO data**

**Exercise materials and tasks**

* **Exercise 5.4 EDA Satellite Imagery using EarthPy** (Group Work)  
   In this exercise, you will learn cloud masking.   
  First, download the notebook from Moodle and upload it to your Google Drive. Then, open the notebook. Read all the instructions, comments and notes within the notebook thoroughly and work through the codes. There are 2 questions and four discussion question. The discussion question does not require a submission.  
   **Question 5.4.1 and 5.4.2:** You will be given a paper with the questions. The questions are also part of the notebook. Please discuss with your group mates, write your answer on the paper, and submit the paper.
* **Code along: EDA on Landsat data using Earthpy**

**Application exercise: GUI-based visualization of EO data**

1. **Familiarisation with the EO Browser from Sentinel Hub**

* Open up the EO Browser from Sentinel Hub
* Direct to Rwanda as your area of interest
* Familiarise yourself with the different índices and take note of the appearance of water bodies, residential areas, forests etc.
* Please provide your insights on the result?

1. **Using the Custom Script repository**
   1. Open a new tab and type “*Sentinel 2 SAVI”*
   2. Open the first option: [Script Repository](https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/savi/%23)
   3. Click on “*copy script”* and add it to your EO Browser
2. **Comparing of indices and time** 
   1. Compare two images from the rainy and the dry season using the NDVI
   2. Compare the NDVI and NDWI from the same image

1. **(if time permits) Visualising forest fires with the NBI**
   1. Fire Occasion in ….
   2. From the Custom Script respository find the script for the NBR
   3. Select the time range to a time after the fire
   4. Add this to your ‘repository’ for comparison
   5. Select the time range to a time before the fire

Time for group work: 30 min

**Videos or Quiz questions**

* A Hitchhikers guide to Hyperspectral Data (8:32) <https://youtu.be/gtlJetaKi_A?t=2447>
* [How to visualise Satellite Images in EO Browser](https://www.youtube.com/watch?v=7Tec2-FzUZQ&pp=ygUYc2VudGluZWwgMiB2aXN1YWxpc2F0aW9u)

**Handouts**

Refer to separate Powerpoint presentation: Visualization of EO data