CIS 4496 - Projects in Data Science

EY Open Science Data Challenge
Spring 2024

1. Introduction:

Navigating the aftermath of natural disasters presents a critical challenge: the swift and accurate assessment of infrastructure damage. Traditional damage assessment methods, which lean heavily on ground surveys and manual checks, are both overwhelming and time consuming. This leads to delayed aid, financial losses, and prolonged suffering for affected communities.

Responding to this challenge, our project proposes a solution rooted in the application of machine learning to analyze satellite imagery for damage assessment. This initiative aims to significantly enhance the speed and accuracy of evaluations post-disaster, thereby enabling a more effective allocation of resources and support to affected communities. Transitioning from ground-based to satellite-based assessments, we leverage advanced object detection and classification techniques to identify damaged infrastructure. This methodology not only ensures a rapid and reliable assessment but also a comprehensive analysis that ground surveys might miss, thereby providing a crucial tool for disaster response teams, insurance agencies, and governmental bodies. These stakeholders, who are in dire need of quick and accurate data, can thus make informed decisions swiftly, ensuring that resources and support are allocated efficiently.

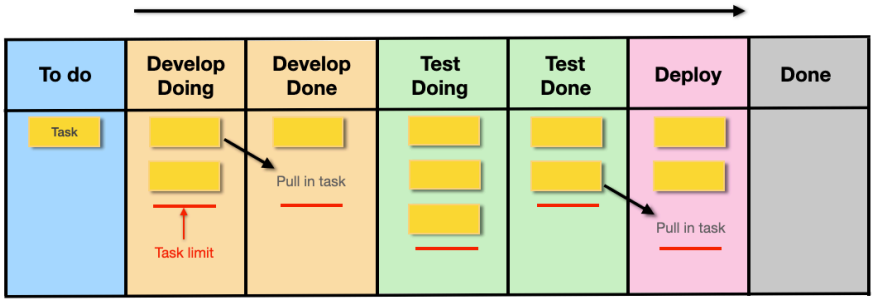
Our approach combines high-resolution satellite imagery, provided by EY, with cutting edge machine learning algorithms. This strategy is particularly underscored by our focus on San Juan, Puerto Rico, where the devastating impact of Hurricane Maria in 2017 highlights the urgency of preparedness for improved disaster response strategies. With our targeted approach, we aim to refine our model's capabilities, specifically for scenarios where hurricanes strike with little to no warning.

The innovation of our project lies in its capacity to transform the landscape of disaster assessment. Traditional methods, while valuable, fall short in the face of large-scale disasters where time and access are limited. Our machine learning model, by contrast, is designed to provide rapid, accurate, and comprehensive assessments, revolutionizing how we respond to natural disasters. This capability is crucial for disaster response organizations, insurance companies, and government agencies, operating in realms where immediate and precise damage evaluation is essential for expediting recovery, processing insurance claims, and optimizing resource allocation.

2. Personnel:

- Team roles:
    - Will Schenk: Team Lead/Project Manager, Feature Engineer
    - Mary Le: Data Quality Engineer, Product Manager

    - Elle Nguyen: Feature Engineer, Model Researcher
- Clients:
    - Business client - EY (Ernst & Young):
        - A multinational professional company provides consulting, assurance, tax and transaction services that help solve clients' toughest challenges.
        - EY can utilize our robust machine learning model to automate the assessment of property damage, facilitating more informed decision-making regarding asset portfolios and investments.
        - Accurate detection of damaged and undamaged buildings post-cyclone can help streamline EY's insurance claims process.
    - Federal client - FEMA (Federal Emergency Management Agency):
        - An agency promotes disaster mitigation and readiness and coordinates response and recovery following the declaration of a major disaster.
        - FEMA can utilize our working machine learning model for rapid damage assessment and resource allocation during and after natural disasters.
        - By automating the identification of affected areas and prioritizing response efforts, responders can allocate personnel and supplies more effectively, potentially saving lives and reducing recovery time.
- Communications:
    - Two in-person weekly meetings to discuss progress and address any concerns with external mentors.
    - Dedicated internal communication channels on Discord and joint review sessions at key milestones (e.g., after model selection, before evaluation) with external mentors.
- Workflow management:
    - Utilizing GitHub as the central platform for code management and documentation.
    - Kanban principles will improve productivity (To-do, In progress, Done)
        - Tasks will be placed in the *Issue Section* of our GitHub repository.
        - Each task will be assigned to a team member and progress will be routinely updated.



3. Problem Description:

a. Project Scope:

Objective

Detect 4 different objects in a satellite image of a cyclone impacted area (Undamaged residential building, Damaged residential building, Undamaged commercial building, Damaged commercial building) using bounding boxes with various object detection models.

Model Infrastructure

Train a supervised machine learning model using the open source YOLO *You Only Look Once*. Specifically, YOLOV8 is state of the art, has strong community support, allows for training on different data formats, and compatibility with other infrastructure like RoboFlow and the XView Spacenet Project.

Training Data

Our training data consists of satellite imagery, paired with polygon segmentation and bounding box detection annotations. We have a variety of resources differing in data type and quality, but the data related to the island of San Juan and Puerto Rico, will serve as a priority in development and testing of our model.

Below is a list of our dataset resources:

- **San Juan City Satellite Imagery**
  - **About:** The San Juan, PR dataset, provided by EY, is satellite imagery of before and after Hurricane Maria in 2017.
  - **Format:** 2 Large TIFF files of before and after, along with post event shapefile annotations
  - **Metrics:** Both files are each 40,000 x 75,000 pixels and each pixel covers 0.305^2 meters. That is 280,000^2 kilometers.
  - **Annotations:** The additional “building footprint” mask includes shapefiles of bounding boxes of medium quality. These identify the single class instances of buildings in San Juan ~ 100 MB. This set of annotations is only present for the post event imagery.
  - **Preprocessing:** Prior to training, python scripting will convert the large TIF images into smaller JPEG formatted pictures known as tiles, and shape file annotations into text file bounding boxes in YOLO format.

- **Rio De Janeiro City Satellite Imagery**
  - **About:** The open source Spacenet Database by Maxar Technologies offers satellite imagery annotated with single class instances of building bounding boxes from Rio De Janeiro, Brazil in 2017. Rio De Janeiro's similar geography as a

coastal city containing forests, mountains, and densely populated urban areas makes this a great addition for building training data.
    - **Format:** TIF images paired with GEOJSON object segmentation annotations
    - **Metrics:** 6940 images totalling 3.1 GB in size, each spanning 200 x 200 m, with over 1 million segmentations of buildings in total.
    - **Annotations:** There are polygon segmentation annotations of medium quality paired with each image.
    - **Preprocessing:** Prior to training, python scripting will convert TIFF to JPEG images known as tiles, and the GEOJSON annotations into YOLO segmentation format.

- **RoboFlow 4-Class Satellite Imagery**
    - **About:** By crowd sourcing information from the RoboFlow website, we were able to obtain 250 unique images of San Juan with high quality detection annotations identifying the building types: undamaged residential, undamaged commercial, damaged residential, and damaged commercial.
    - **Format:** JPEG Images paired with YOLOV8 formatted annotations
    - **Metrics:** (512 x 512) pixels
    - **Annotations:** Annotations are provided in the form of bounding boxes, each defined by a list of 4 coordinates. These coordinates precisely delineate the perimeter of each building within the images. Accompany each bounding box is a label from the set {0, 1, 2, 3}, categorizing the building types.
    - **Preprocessing:** Imported Albumentations and PIL library to apply a series of augmentation techniques include, horizontal and vertical flips, rotations, random brightness and contrast, saturation, and blur. Since the augmentation alters the image layout, the coordinates of each annotation's bounding box are updated to reflect these changes, ensuring that the annotations remain accurate and aligned with their corresponding images.

- **XView Dataset:**
    - **About:** The XView dataset is the well known publicly available satellite imagery object classification dataset and was published by the Department of Defense in 2018. It detects a massive range of objects and its training data spans the globe.
    - **Format:** TIF images paired with GEOJSON object detection annotations
    - **Metrics:** 1,415 km^2 span, 0.3 meter resolution, and 1 million object instances
    - **Annotations:** There are 60 annotation classifications, most notably, trucks, cars, buildings, and sheds.
    - **Preprocessing:** YOLO's website offers a guide to working with the XView data and converting it to the needed format. Through experimentation, training the XView dataset with YOLOV8 and testing with the San Juan data, the change in

zoom of the two datasets lessened its ability to make predictions. Therefore, to adjust for zoom, we will re-process the XView data to match the same zoom depth of the San Juan data, which is approximately 0.305^2 meters per pixel, and 40,000 x 75,000 pixels per image.

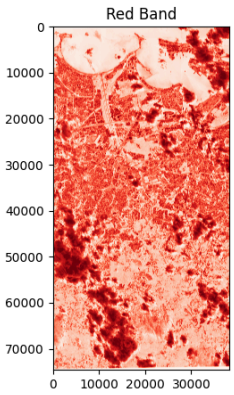
b. Metrics:

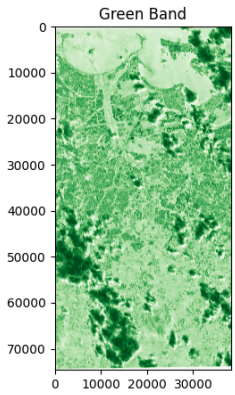
To test our model, we will make predictions for the 4 building classifications on the satellite imagery of San Juan following hurricane Maria. Fortunately, EY provides a dataset of correct annotations laid over 20 images in San Juan after the hurricane. Here, we will calculate mean average precision and consistently work to improve this metric.

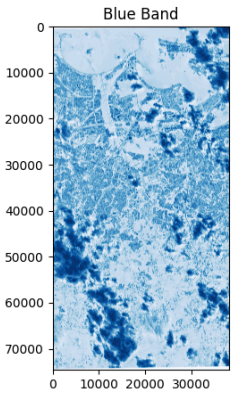
- *We need here specific Computer Vision evaluation metrics such as:*
  - *Precision, Recall;*
  - *Intersection over Union;*
  - *Average Precision, Average Recall used for object detection;*
- Qualitative Objectives: Improve accuracy and effectiveness in storm damage assessment.
- Quantifiable Metric: Reduce the time for initial damage assessment by 50%.
- Desired Improvement: Decrease the time taken for storm damage assessment by 50% compared to the current manual process.
- Baseline Value: The current time taken for storm damage assessment is 2 days.
- Measurement: Time taken for storm damage assessment before and after model optimization will be compared.

c. Resources:

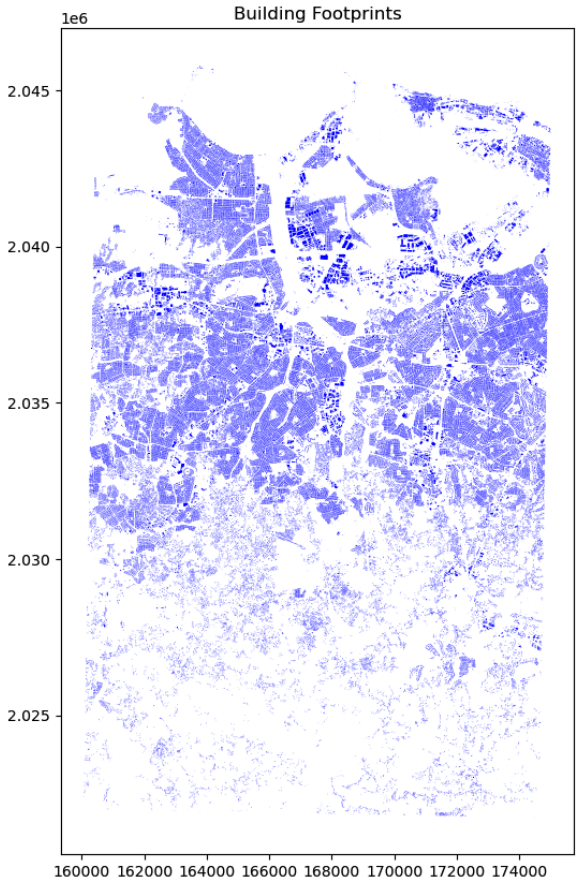
- Annotation tools: Labelme, Labelbox, etc.
- Libraries: Geopositioning and Image Processing Libraries include GDAL, Rasterio, Ultralytics, and GeoPandas.
- Supporting data provided by the competition:
  - Raw pre- and post-cyclone satellite images as TIFF (Tag Image File Format) files provided in the competition's available notebook.
    - Below is a sample of a pre-event imagery in each color decomposition (Red, Green, Blue) with clouds as darker regions.
    - This is obtained by using Python's GDAL utilities to visualize each imagery without the installation of any geographic information system software (for example, QGIS).







- Building footprints



4. <u>Plan:</u>

We will iterate through the data preprocessing stage to ensure all data is ready to be trained with YOLO. We will have at least the following 4 datasets:

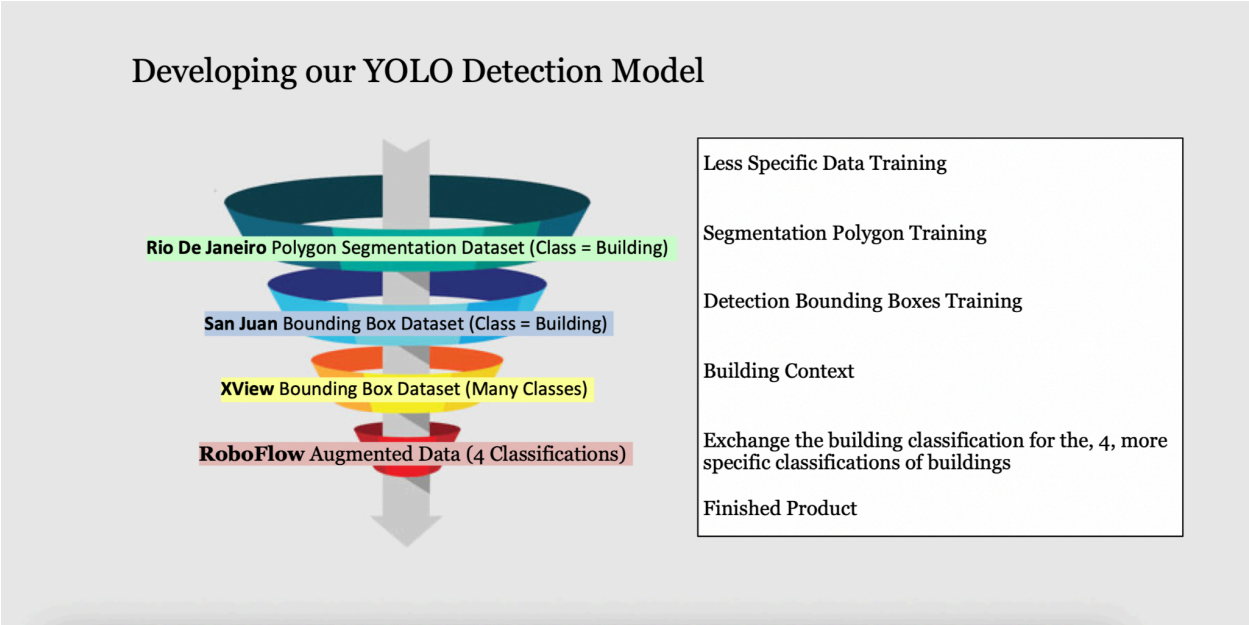
- We have Roboflow Augmented Data (Annotations (4-types of buildings) and Images)
- We have Rio De Janeiro Data (Annotations (buildings) and Images)

- We have San Juan Data (Annotations (buildings) and Images)
- We have the XView Data (Annotations(many objects) and Images)

Additionally, we can process more data, specifically from geological data like terrain, elevation, etc. This will likely be in a segmentation format and come early in the model training process.

Now, we will creatively train our model with transfer learning. We will train and fine-tune the neural network weights until we have reached our end product, which will be a model able to predict the state (damaged or undamaged) and type (residential or commercial) of the buildings.

For example, we can train on one dataset and then use this model's weights and retrain it on another dataset. YOLOV8 specifically allows segmentation annotations to be trained initially prior to detection annotations, so we will begin with the Rio De Janeiro segmentation data. Then, adding San Juan, followed by XView bounding box data. Finally, we will adjust our model settings to create predictions for the four unique classifications with the RoboFlow dataset.



The metrics finally used to evaluate our model are outlined in the Metrics section above.

- By 03/01, we will complete **Phase 1 - Data Collection**
    - Gather high-resolution panchromatic satellite images given by [EY Open Science Data Challenge Platform], including pre- and post-event images of affected areas, along with building footprints.
    - Obtain 2-3 more satellite imagery datasets for model training.

        - Update: we obtained 4 more supporting datasets that will increase object detection as well as building classification performance.
            - With annotated labels of buildings:
                - Rio De Janeiro Data
                - San Juan Data
            - With annotated labels of many objects at once:
                - XView Data
            - With annotated labels of buildings with 4 different classes:
                - Roboflow Augmented Data (See: Project Scope for data augmentation process)
- By 03/15, we will complete **Phase 2 - Data Preprocessing**
    - Instead of going through the manual annotation phase, we will preprocess 4 previously obtained datasets with readily available annotations in bounding boxes format.
    - We created a script for data augmentation using Python libraries such as PIL and Albumentations
- **Phase 3:** Model Development (2 weeks)

    The team employs a machine learning model to detect objects (Damaged residential buildings, Undamaged residential buildings, Damaged commercial buildings, and Undamaged commercial buildings)
    - **Phase 4: Model Evaluation and Improvement**
    - Conduct prototype testing of both ResNet-50 and YOLOv8 on validation dataset to compare performance
    - Optimize training epochs by employing k-fold cross validation to determine model's performance across different subset of the data.
    - Model improvement: address class imbalance by implementing SMOTE or adjusting loss functions to weigh rare classes more heavily.

The model's accuracy and effectiveness are evaluated using a separate set of images. Refinements will be made to improve the model's predictive capabilities. Anticipated areas of model improvement include:

- Create training data
- Optimize class imbalance
- Explore object detection algorithms
- Leverage moderate resolution satellite data (Sentinel-1, Sentinet-2)
- Optimize training approach

5. Challenges:

- **Identifying Instances of Buildings from Satellite Imagery:**

    - **False Positives:** We must avoid false positives of on-ground objects that may appear similar to buildings in satellite imagery. For example, pools or trucks with their large rectangular shape will appear similar to buildings when looking down from above.
    - **Large Variation:** Buildings vary drastically in shape, size, color, and texture.
    - **Obstructions:** Trees, clouds, reflections, and darker shaded regions can obstruct our clear, birds-eye sight of the building.
    - **Unique Buildings:** Deciphering what buildings are unique is vital. For example, closely built structures may appear as single instances of buildings.

- **Zoning Regulations:**
    - While certain areas may appear to be predominantly commercial or residential based on Google Maps view, detailed zoning regulations dictating land use are often inaccessible. Such multifunctional nature of buildings often complicate our classifications, making it challenging to effectively label as either commercial or residential.
    - For instance, a multi-story building might be utilized for residential purposes on most floors, with the ground floor allocated for commercial activities (apartments above a street-level cafe).

- **Model Training Issues:**
    - **Overlap:** The rectangular bounding boxes of different buildings may overlap due to the position of the buildings.
    - **Alignment:** Structures in satellite imagery appear slanted since the satellite is not perfectly aligned over the objects. This can cause the bounding boxes to appear slightly shifted from actual building blocks.
    - **Varying Zoom:** The model may have varying success if the model is trained on one set of images and then tested on new images with a different zoom (far or close) and quality.
    - **Storage:** Since we are training neural networks with large quantities of images, a large volume of storage space is needed. Our primary dataset of pictures of San Juan before and after Hurricane Maria are 1 GB each. After bringing in external datasets and conducting experiments, a total of 25 GB of storage has been needed.
    - **Computing power:** Through our experiments, training detection and segmentation neural network models have occasionally failed without GPU processors or even with free computing powers provided by Google Colab.