



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence

# Data Gathering & Annotation

---

Andrea Filibeto Lucas

# Quick Introduction...



## What is Data Gathering?

The process of **collecting raw information** from various sources e.g., cameras, sensors or online datasets)

In AI projects, this step provides the **foundation for model training**.

## What is Data Annotation?

The process of **labelling collected data to give it meaning and structure** for machine learning.

Each data sample is **tagged with relevant information** (e.g., bounding boxes, categories, or attributes).

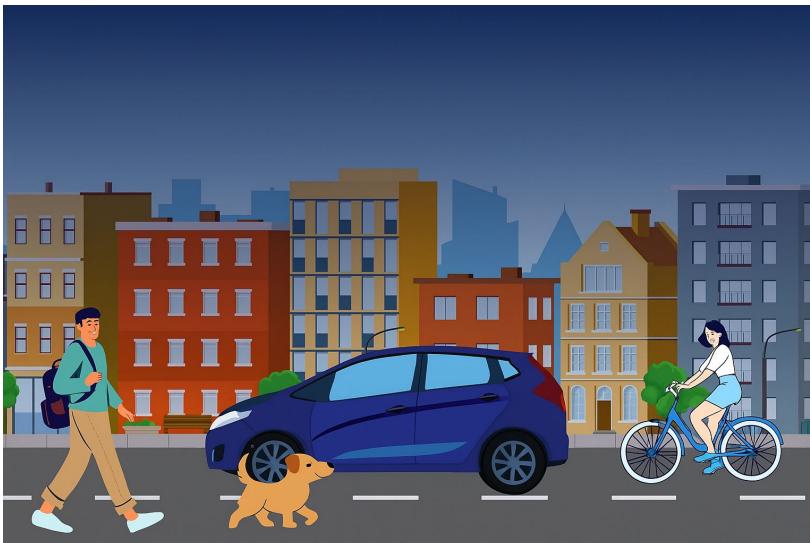
# Or More Simply...



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

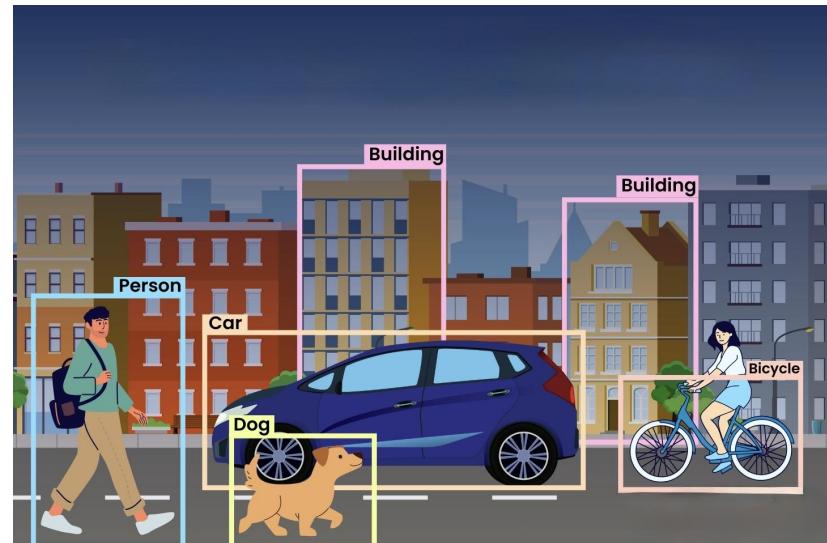
Department  
of Artificial  
Intelligence

## Data Gathering



Sourced from: [www.projectpro.io/article/data-annotation-in-ai/1154](http://www.projectpro.io/article/data-annotation-in-ai/1154)

## Data Annotation



NB: This is an incomplete annotation - not all buildings/people are annotated.

# Types of Annotation



## 1. Image Description

Provides a textual summary of the image.

Common in captioning datasets (e.g., “A red stop sign on a rural road”).

## 2. Box Annotation

Uses bounding boxes to mark specific objects.

Essential for object detection tasks (e.g., locating traffic signs).

## 3. Segmentation

Divides the image into pixel-level regions.

Used for precise object boundaries (e.g., segmenting a road sign’s exact shape).

## 4. Keypoint & Landmark Annotation

Marks specific points of interest (e.g., facial features, joints, or pole mounting points).

# Good vs Bad Annotation



Bounding boxes **TIGHTLY FIT** around objects.

Correct labels are applied.

**Consistent style across images** (same criteria, naming, and labelling precision).

All visible instances are annotated - none are missed.

## Why This Matters?

Poor annotation leads to misleading training data, causing models to learn incorrect patterns and perform unreliably in real-world conditions.



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence

# Annotation platforms

---

# What are Annotation Platforms?

Annotation platforms are **software tools** designed to label and manage datasets for ML.

They allow users to **draw boxes**, masks, or keypoints on images, assign class labels, and **export data in standard formats**.

These platforms streamline the process of preparing high-quality, structured data for training CV models.

Examples:

- [Label Studio](#), CVAT (CV Annotation Tool) or Roboflow

NB: For your assignment, you must use Label Studio.



L-Università ta' Malta  
Faculty of Information & Communication Technology

Department of Artificial Intelligence



# Export Annotation Formats?

## COCO Format

- Stores *images*, *annotations*, and *categories* in one file.
- Used by YOLOv8, TensorFlow, etc.

## YOLO Format

- One `.txt` per image: `class x_center y_center width height`.
- Compact, normalised (0–1), ideal for real-time detection.

## Label Studio / JSON

- Keeps full metadata and attributes.

## Conversion

- Libraries like **Ultralytics** or **custom scripts** convert between formats.
- **Always check coordinate systems after conversion.**



```
1  {
2      "id": 3,
3      "image_id": 2,
4      "category_id": 1,
5      "bbox": [
6          124.06,
7              86.66,
8              90.61,
9              85.87
10         ],
11         "area": 7780.68,
12         "iscrowd": 0,
13         "attributes": {
14             "view_angle": "Front",
15             "mounting": "Pole-mounted",
16             "condition": "Good",
17             "sign_shape": "Circular"
18         }
19     }
```

(COCO Annotation Sample)



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence

# Gathering Good Data

---

# Good vs Bad Data Gathering



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

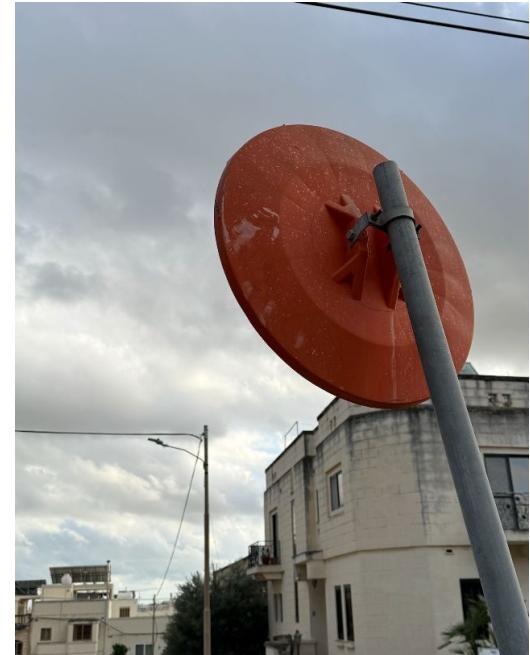
Department  
of Artificial  
Intelligence



*Front*



*Side*



*Back*

# Good vs Bad Data Gathering



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

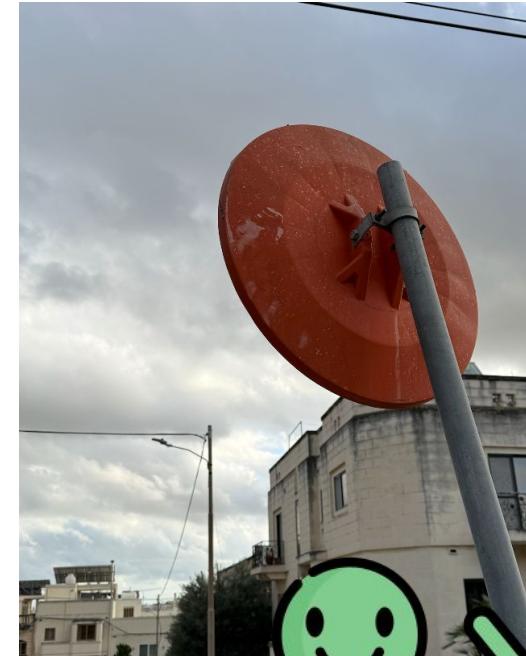
Department  
of Artificial  
Intelligence



Front



Side



Back

# Case: Sign not installed correctly



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence



*Front*



*Side*



*Back*

# Case: Sign not installed correctly



Front



Side



Back

# Case: Sign not installed correctly



L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence



Front



Side



Back



**Poor data gathering, such as capturing visible number plates or unclear sign views, violates GDPR standards and reduces dataset quality. Signs should be clearly visible (even if not centred).**



# Setting-Up Label Studio

---

Please follow the demonstration.

# Getting Started: Installation



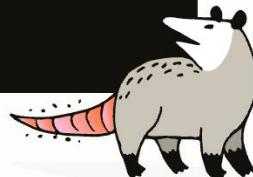
L-Università ta' Malta  
Faculty of Information &  
Communication Technology

Department  
of Artificial  
Intelligence

## Quick Start

```
1 # Install the package
# into python virtual environment
2 pip install -U label-studio
3
4 # Launch it!
5 label-studio
```

PIP BREW GIT DOCKER



## Quick Start

```
1 # Install the cask
2 brew install humansignal/tap/label-studio
3
4 # Launch it!
5 label-studio
```

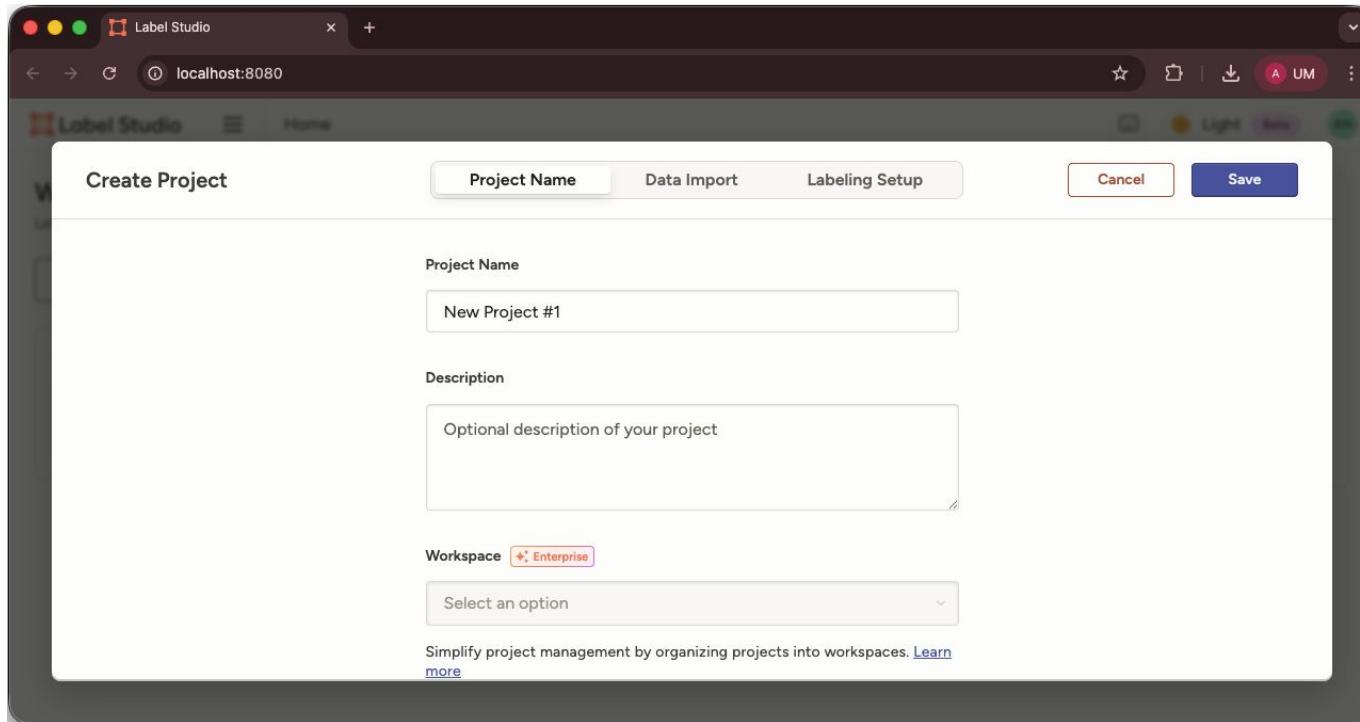
PIP BREW GIT DOCKER



*Ensure you have Python (up to version 3.12) & pip/brew installed on your system.*

*For more information visit: <https://labelstud.io/>*

# Step 1: Create a New Project



After starting Label Studio (at <http://localhost:8080> - by default), click "Create Project".

# Step 2: Import Your Data



The screenshot shows the 'Create Project' interface in Label Studio. The 'Data Import' tab is active. A central area allows users to 'Drag & drop files here or click to browse'. Below this, a list of supported file types is provided:

Images	bmp, gif, jpg, jpeg, png, svg, webp
Audio	wav, mp3, flac, m4a, ogg
Video ⓘ	mp4, webm
HTML / HyperText	html, htm, xml
Text	txt

You can upload files (images, text, CSVs) directly from your computer, or you can connect to cloud storage.

# Step 3: Set Up Labeling Interface



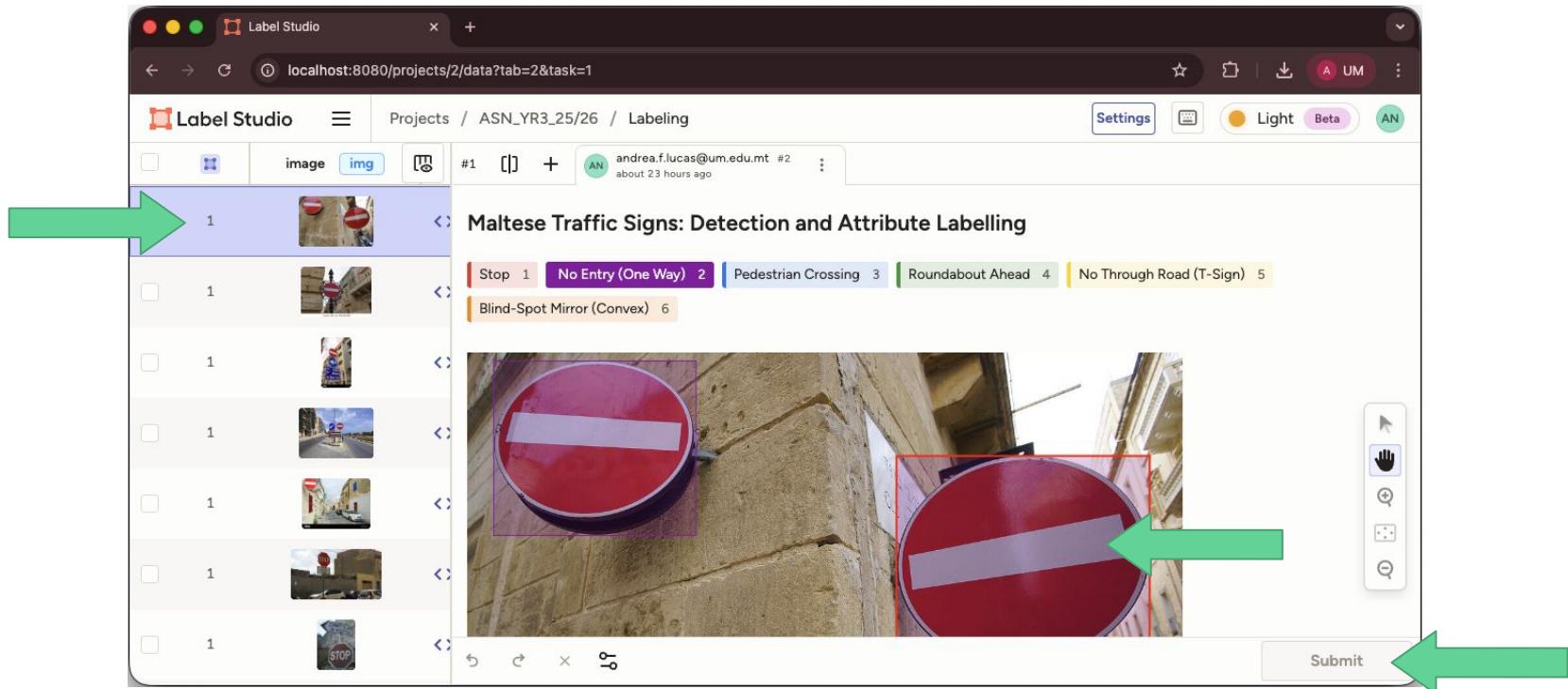
The screenshot shows the Label Studio interface for setting up a labeling interface. On the left, a sidebar lists options like General, Labeling Interface (which is selected), Annotation, Model, Predictions, Cloud Storage, Webhooks, and Danger Zone. The main area has tabs for 'Browse Templates', 'Code', and 'Visual'. A large green arrow points from the 'Code' tab to the XML code editor. The XML code is as follows:

```
1 <View>
2   <Header value="Maltese Traffic Signs: Detection and Attribut...
3   ...
4   <!-- OBJECT -->
5   <View style="display:flex;align-items:start;gap:8px;flex-wrap:wrap">
6     <Image name="image" value="$image" zoom="true" zoomControl="false" ...
7     <RectangleLabels name="sign_type" toName="image" show="true" ...
8       <Label value="Stop" category="1" background="#E53935" ...
9       <Label value="No Entry (One Way)" category="2" background="#F0A0D2" ...
10      <Label value="Pedestrian Crossing" category="3" background="#A0C0F0" ...
11      <Label value="Roundabout Ahead" category="4" background="#FFCCBC" ...
12      <Label value="No Through Road (T-Sign)" category="5" background="#FFB74D" ...
13      <Label value="Blind-Spot Mirror (Convex)" category="6" background="#FFC107" ...
14    </RectangleLabels>
15  </View>
16
```

Below the code editor, there's a note: "Configure the labeling interface with tags. See all available tags." To the right, a list of available tags is shown with their counts: Stop (1), No Entry (One Way) (2), Pedestrian Crossing (3), Roundabout Ahead (4), No Through Road (T-Sign) (5), and Blind-Spot Mirror (Convex) (6). Further right is a visual interface showing an aerial view of an airport tarmac with several airplanes, overlaid with red bounding boxes and labels. A legend on the right side of the visual interface shows icons for selection, zoom, and other tools.

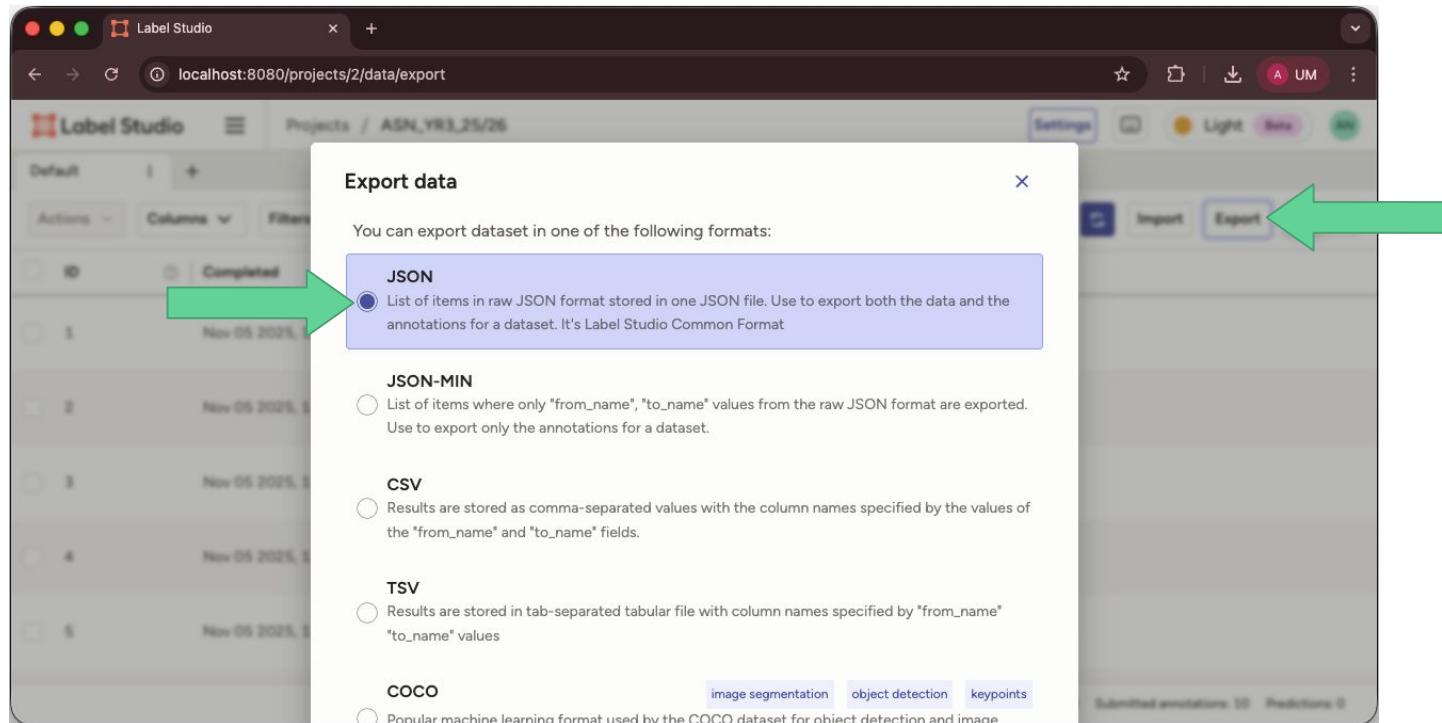
*For your assignment, the XML file will be provided to you.*

# Step 4: Start Labeling



Click "Label All Tasks" from the project dashboard begin. Select your labels (e.g., click "No Entry") and press "Submit".

# Step 5a: Export Your Annotated Data



Once you've labeled your data, click the "Export" button on the project page & choose your desired format.

# Step 5b: Export Your Annotated Data



The screenshot shows the Label Studio export interface at <localhost:8080/projects/2/data/export>. A green arrow points to the 'COCO with Images' option, which is highlighted with a blue background. Other options like 'JSON', 'JSON-MIN', 'CSV', 'TSV', and 'Pascal VOC XML' are also listed.

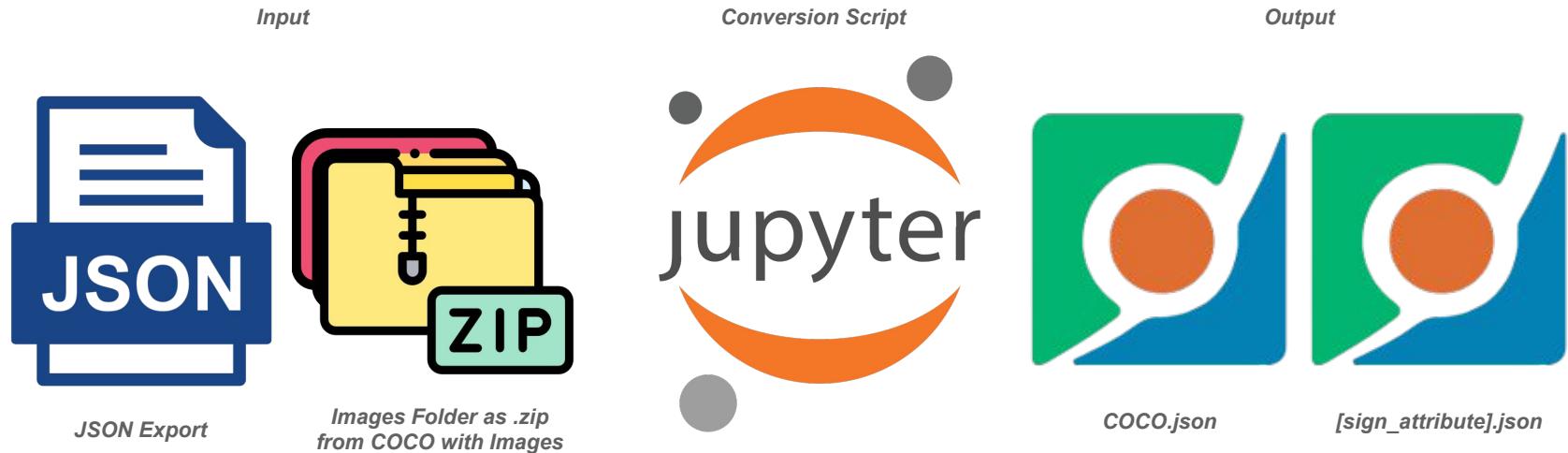
- List of items in raw JSON format stored in one JSON file. Use to export both the data and the annotations for a dataset. It's Label Studio Common Format
- List of items where only "from\_name", "to\_name" values from the raw JSON format are exported. Use to export only the annotations for a dataset.
- Results are stored as comma-separated values with the column names specified by the values of the "from\_name" and "to\_name" fields.
- Results are stored in tab-separated tabular file with column names specified by "from\_name" "to\_name" values
- Popular machine learning format used by the COCO dataset for object detection and image segmentation tasks with polygons and rectangles.  
COCO with Images image segmentation object detection keypoints  
COCO format with images downloaded.
- image segmentation object detection

*Scroll down and export using 'COCO with Images' to get the updated image filenames.*



**Important for your assignment: you need JUST the images folder from the COCO (with images) export and the JSON file from the Label Studio JSON export. Please follow the provided notebook carefully for the conversion to work correctly.**

# Step 6: Convert Annotations



**NB:** Use the provided Jupyter notebook with the JSON annotations and corresponding images (as a .zip file) to generate the assignment's COCO-format annotations.



# Any Questions?

 Andrea Filiberto Lucas



[andrea.f.lucas@um.edu.mt](mailto:andrea.f.lucas@um.edu.mt)

 Dylan Seychell



[dylan.seychell@um.edu.mt](mailto:dylan.seychell@um.edu.mt)

# Further Reading & Resources



- <https://labelstud.io/>
- <https://roboflow.com/>
- <https://www.cvat.ai/>
- <https://www.youtube.com/watch?v=R1ozTMrujOE> (*Label Studio Tutorial*)
- <https://blog.roboflow.com/how-to-use-label-studio/>
- <https://labelstud.io/learn/getting-started-with-label-studio/>
- <https://blog.roboflow.com/yolov11-how-to-train-custom-data/>