The dataset sourced from Aicrowd had 414,424 photographs belonging to 772 snake species and taken in 188 countries. The majority of the data were gathered from online biodiversity platforms (i.e.,iNaturalist, HerpMapper) and were further extended with noisy data scraped from Flickr containing wrong labels and out-of-scope (non-snake) images.

**Challenges faced**

A number of challenges were faced by Myself and David this week to prepare our dataset for training the model. Upon downloading the dataset(<https://www.aicrowd.com/challenges/snakeclef2021-snake-species-identification-challenge/dataset_files>), we realised that:

1. None of the images were organised into files
2. All images were random / different sizes
3. There was a different number of images per snake
4. Downloading the dataset from AiCrowd

Dowloading the dataset was a challenge because is size was around 60GB and was hosted on a S3 AWS in America which made the file really hard to download at a decent speed, but even after 1 day when it reached 30GB the file was failing because of authentication AWS expiry problem.

I had to install a VPN and change my location to America, only then I was able to download the dataset.

Once I received the text files with the right image snake name that David prepared I had to do a batch copy of the snake images into the right folder to assign a label and prepare the dataset.

For doing this I have used the command : FOR /F "delims=" %N in (list.txt) do COPY "%N" Target

* 'for' loops through all the filenames in list.txt
* "delims=" says use the whole of each line as the parameter %N
* the quotes around %N in 'copy "%N"' allows for any filenames that contain spaces
* Target specifies the folder you want to copy to

So I was able to copy the files list created from David into specific folders for creating the labels and dataset.

After I went into each folder and observed the data , found a couple of image not that were not snakes and removed them , then resized the dataset to 490pic per snake except one that is 425 and uploaded to my drive for everyone to see and download:

<https://drive.google.com/drive/folders/197gu68UTmiYhGvvDVHe9YZwhfvu_qfhJ?usp=sharing>