**Plan of Action**

We will be using Image form of non-structured data (Lung Xray images) for this project

Data Source: <https://www.kaggle.com/datasets/fatemehmehrparvar/lung-disease>

For our data pipeline we will follow ML/AI project lifecycle:

Step 1: Data Downloaded from data source

Step 2: Analyze and preprocess the data

* Remove duplicate files (check by file name) (no duplicates as lungs are already numbered in numbers)
* For data uniformity resize all images
* Normalization of pixel values
* Apply rotation, zoom, brightness augmentation if needed or make images grayscale
* Split data into train and test and save them in preprocessed folder

Step 3: Model Development and Evaluation

* Choose models based on Step 2 findings (most probably will be Neural Network models)
* Train model
* Evaluate metrics used to evaluate Classification models (e.g. Accuracy, Precision, Recall )