

# Project Proposal



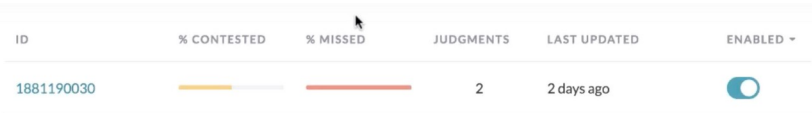

Mohammed Almuylibi

---

## Data Labeling Approach

<b>Project Overview and Goal</b>  What is the industry problem you are trying to solve? Why use ML in solving this task?	<b>The problem here is a problem of categorization of x-ray images based on the presense of pneumonia's signs in the images. ML can categorize and detect detials faster and better than human since it turns images and its features into numbers.</b>
<b>Choice of Data Labels</b>  What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	<b>Pneumonia, healthy, unknown , others.</b> <b>Pneumonia means the Pneumonia signs are present in the image.</b> <b>Healthy means the image shows no Pneumonia signs are not present in the image.</b> <b>Unkown means there are signs of abnormalities in the image but is not like Pneumonia.</b> <b>Other means there are signs of abnormalities but is not Pneumonia.</b>  <b>I chose these labels to ensure that any new data will fit into one of the cataegories. Pneumonia is set already while Other will be leading to other categories depends on the incoming data.</b>

## Test Questions & Quality Assurance

<p><b>Number of Test Questions</b></p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>	<p><b>I've created 12 questions</b></p>
<p><b>Improving a Test Question</b></p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>	 <p><b>I believe in this case you need to redesign the job and change the instructions to be clear.</b></p>
<p><b>Contributor Satisfaction</b></p> <p>Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)</p>	 <p><b>We need to add more data and fouces more on the rules.</b></p>

# Limitations & Improvements

<b>Data Source</b>  Consider the size and source of your data; what biases are built into the data and how might the data be improved?	<b>The dataset is small and I haven't seen any biases till now. Biases are an issue that happen unknowingly as it is impact the performance and the quality of its result. To ensure that there is no biases in the data we have to monitor the model performance, by using data with known label for us to ensure the quality of our model.</b>
<b>Designing for Longevity</b>  How might you improve your data labeling job, test questions, or product in the long-term?	<b>We need to add more data on a regular basis to ensure we have a large dataset that contains many examples with different labels.</b>