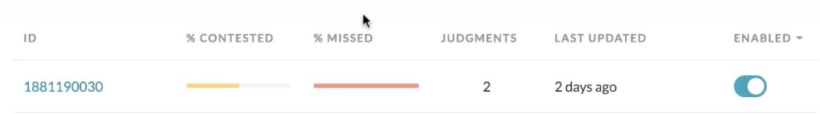
Project Proposal

Mohamed Amr

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	I am trying to build a ML system for the healthcare industry to aid in quickly identifying healthy patients and surfacing potential cases of pneumonia.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	Basically, I added two main labels that annotators could make their decision on. The first label is for **pneumonia** images, and the second for **normal** images. Moreover, I added a third label (**Not Sure**) to guide annotators. I ask annotators to describe how likely they think a case of pneumonia is in a given image, by rating it from 1 to 5.

Test Questions & Quality Assurance

Number of Test Questions Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?	I will consider the ground truth data. So, I will add sixteen test questions.
Improving a Test Question Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	 I may need to augment the instruction or include more examples or redesign the job.
Contributor Satisfaction 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.)	 I will add more reasons on the test questions, add more tricky cases in the examples, and I will add more examples for each possible data annotation. Also, I will improve the steps and instructions to make the annotation job more clear.

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The original dataset contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images and 2 categories (Pneumonia/Normal). It's biased for children X-Ray. So, we can improve it by collecting more X-Ray images for different age groups. Moreover, we may consider gender effect, and we can add more labels to the images with (Male, Female, Other .. etc.)
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	I can improve data labeling job by designing for failures and longevity. As I mentioned above, we can improve it by collecting more X-Ray images which is labeled for different age groups and genders.