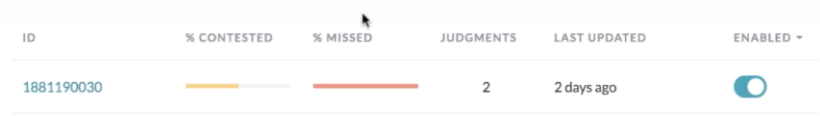
Project Proposal

Amal Aljabri

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Usually difficult for doctors to quickly identify cases of pneumonia in x-ray images. I am trying to build a product that helps doctors quickly identify cases of pneumonia in children. this can be used by ML engineers later on down the line to build a classification product.
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?	I added three labels: • The first label is "Yes" for cases of pneumonia images. • The second label is "No" for cases of normal images. • The third label is "Unknown" for cases of uncertainty.

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 prepared 12 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?	 We may need to the following: • Increasing the instruction. • Addition more examples. • 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.)	 We will need to the following: • Improve on the steps, rules and tips sections. • Addition more examples use cases.

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 dataset is different in size and exposure times. I haven't seen any bias yet. But if there are biases in the dataset, then you need to augment the dataset.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Can be improved as come across new data and with more cases.