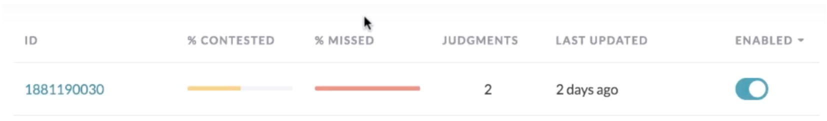
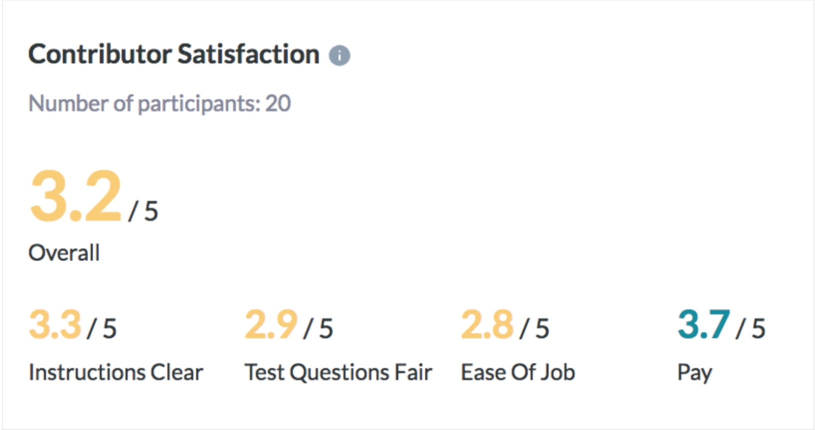
Project Proposal

Rosemary Nwosu-Ihueze

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Pneumonia is a respiratory disease that hospitalizes more than 500,000 Americans with a high mortality rate. This disease cuts across all age demographics with infants and the elderly being high risks. Despite all, this disease can be treated and can be prevented using vaccines. In this project, I would be annotating the x-ray images for easier identification aiding precision medicine for healthcare practitioners. Classification method allows for easy and quick automation of x-ray analysis and with the fast evolution of machine learning in precision medicine, it's accuracy has surpassed manual detection.
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 used binary labelling with yes as present and no as absent. I made an inclusion of a third label which was unknown with the consideration of a non-trained testers and for mild cases or presence of false positives. I chose this method as the objective is to find out if there is presence or absence of the disease which is a classification problem.

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 provided 25 questions to mitigate classification challenges as this will give testers an avenue to evaluate unclear cases and improve confidence levels.
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?	 ID | % CONTESTED | % MISSED | JUDGMENTS | LAST UPDATED | ENABLED ▾ 1881190030 | | | 2 | 2 days ago | ● The high failure rates indicates that either the instructions weren't followed, or the testers were unsure of images and kept choosing the unsure option. To improve this issue is by making the question clearer with more examples of the unknown samples.
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.)	 Contributor Satisfaction Number of participants: 20 3.2 / 5 Overall 3.3 / 5 Instructions Clear 2.9 / 5 Test Questions Fair 2.8 / 5 Ease Of Job 3.7 / 5 Pay Overall rating is poor. Ease of job is challenging to be reduced. To augment, the fairness in questions would be improved by providing more examples.

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 effects in the images such as brightness and opacity can affect the accuracy as some healthy lungs may appear cloudy. To improve this, future images would be taken using devices with same specifications. In a situation there are images taken with devices of varying specifications, an offset would be done. There are different biases that may come into play which includes: • Human Bias: unconscious human bias which is introduced, and this can be improved by acknowledging those biases and account for them. • Algorithm Bias: This can be from the way the model was developed or trained and this can be improved using A/B testing. • Exclusion Bias: This bias is from removal or addition of features during preprocessing when weighing irrelevant/relevant scale. This can be improved by including those with domain knowledge in the analysis and engineering process.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Inclusion of researchers, medical practitioners in labelling. A user study can be run on instruction and example quality. Also based on the feedback of current set of examples, rules and Tips, Test questions