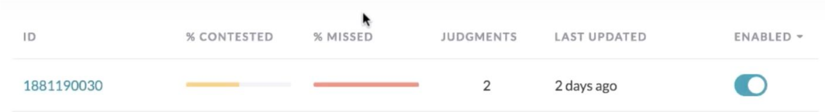
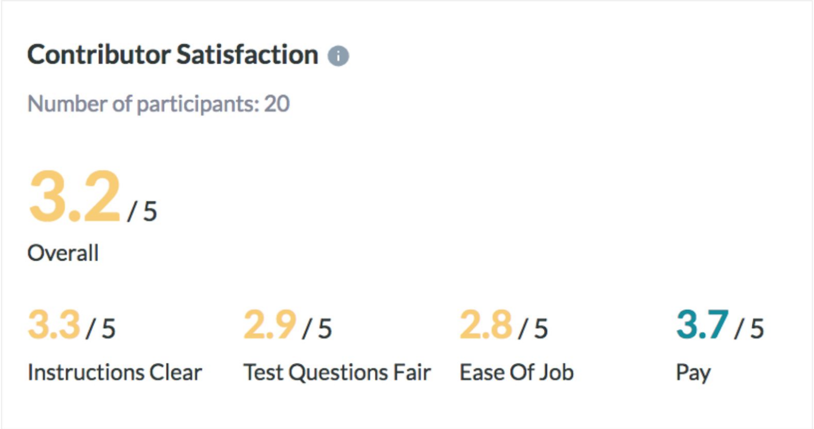
Project Proposal

Bertrand Ndakena

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	**Because of uncertainties in chest x-ray images, it is usually difficult for doctors to quickly identify cases of** pneumonia in those images. This image annotation job is aimed at building a product that helps doctors quickly identify cases of pneumonia in children. **Machine learning(ML)** will then be used to create a classification tool that can automatically tag a chest x-ray image either as **normal**(healthy) or **pneumonia.** Identifying cases of **pneumonia** from images is a classification problem and can best be handled using **ML,** it can offer an objective opinion to improve efficiency, reliability, and accuracy in quick identification of cases of **pneumonia** in children.
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 decided to label the data with **Yes**(meaning presence of pneumonia) or **No**(Meaning normal image or absence of pneumonia) or **Unknown** (for cases of uncertainty). This is a binary classification. Since the aim is to determine if a patient(x-ray image) has pneumonia or not, the best option is to use binary classification, I think it is not reason enough to talk of uncertainty when while trying to determine the presence of a disease, which is the reason why I didn't use likelihood(measuring on the scale 0-n). Also, I would have asked annotators to indicate signs of pneumonia on those images, however, this would further complicate the ML classification. A weakness here is that an annotator won't be able to express his level of certainty and or 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 5 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?	 ID % CONTESTED % MISSED JUDGMENTS LAST UPDATED ENABLED 1881190030 2 2 days ago If many annotators are missing a particular question, I may consider 1. Modifying the instructions to include the question as an example and giving a clear and a more detailed explanation. 2. Adding more test questions to include more difficult cases, it might be reasonable to work with a medical expert here OR 3. Redesigning the entire 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.)	 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 For poor ratings below 3.5, I will have to give a clearer definition of pneumonia, improve on the rules section of my instructions and giving more example 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?	- Data set has images with black particles that are not defined, this can really make it difficult for annotators to make a decision on such images, resulting in many **unknown** labels. If there are many unknown labels, the ML classification may face accuracy issues. It is important to clearly explain what those particles represent, whether there is any relationship between them and **pneumonia**, this can improve the data labeling process. - pneumonia has been defined on the basis of cloudiness/opacity of the lungs and the diaphragm areas, however, this cloudiness could be caused by some other illness different from pneumonia. Also, cloudiness and opacity are not the only signs of **pneumonia**, it may be necessary to include other signs like white spots on the lungs.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	In the long term for changing data, I might have to - Modifying the instructions - Use more relevant examples that are standard, including many difficult cases - Adding more test questions to help annotators understand better