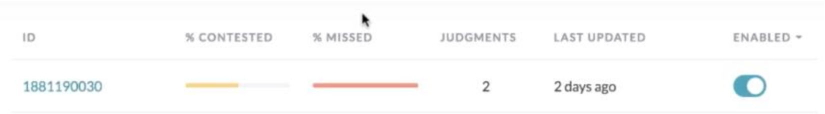
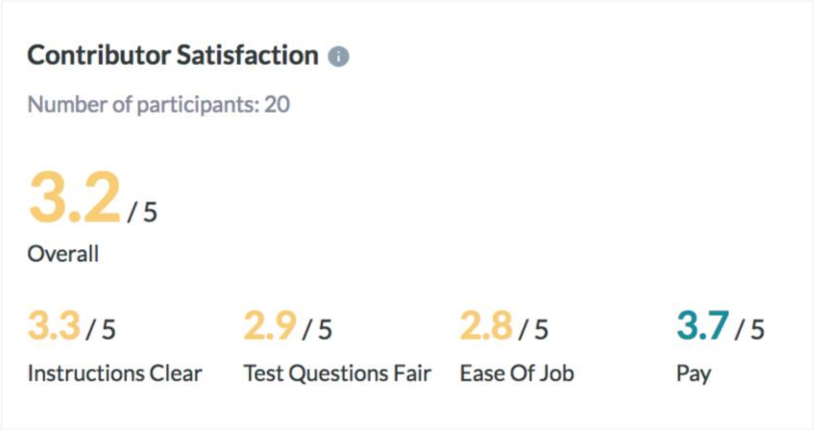
Project Proposal

Vinícius da Silva Vale

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 an infection that inflames the air sacs in one or both lungs. According to Dadonaite B. & Roser M. (2019) in Pneumonia more than 2 million people died from pneumonia in 2017. Ideally, pneumonia would always be diagnosed by a physician using radiological imaging and determining the infectious agent that caused the disease. However, because such diagnosis requires a lot of resources, it is in many cases not done. The purpose of this project is to develop a machine learning model that can identify and distinguish the difference between x-ray of healthy lungs or those with pneumonia to help doctor quickly identify cases of pneumonia in children. Possible benefits will be to decrease the diagnosis time by the doctor, reducing the use of time resources, increasing the productivity of the same, which means greater opportunity for more lives to be saved and also decreasing the likelihood of wrong, undiagnosed or late diagnoses cases. "Diagnostic errors contribute to approximately 10 percent of patients deaths", according to The Institute of Medicine at the National Academies of Science, Engineering and Medicine 2015 report. A study from Beth Israel Deaconess Medical Centre and Harvard Medical Schools (hms.harvard.edu/news/better-together) demonstrated that a Machine learning model trained with images labeled with regions showing cancerous and noncancerous cells achieved a diagnostic sucess rate of 92 percent. Combined with the expertise of a doctor, the success rate rose to 99.5 percent. Wich is much more accurate than 96% achieved by an expert. This project will be divided into three phases: 1. Create a labeled Dataset with images and identifiers of healthy lungs or pneumonia. As it was not possible to hire a specialist capable of labeling the x-ray images, some easily recognizable patterns will be used to allow ordinary people to classify the images. For this purpose, the Appen platform will be used, we will present the cases and guidelines in an interactive questionnaire. The other items in this document provide more information about this phase. 2. Build a Classification model 3. Evaluate the machine learning model
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?	Labels present in the dataset: 0 – Normal 1- Pneumonia As the images will be labeled by non-experts, uncertainty labels will be added: 2 - Unknown 3 – uncertain, maybe Normal 4 – uncertain, maybe Pneumonia The adoption of a 5-point scale allows to deal with uncertainty and the chances of results that are only unknown.

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?	The Pareto Principle, or the 80/20 rule, states that for many phenomena 80% of the result comes from 20% of the effort. As we have 117 cases, 20% represents 23 cases. Of which 16 will be labeled cases and 4 cases of sick and 3 healthy lungs.
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 · First analyze which class belongs to the question. Then use the case as an example or highlight the feature that is causing confusion. In addition, it would include the reason for the error in the question so that note takers can learn and add more similar cases.
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 Identify which questions resulted in the most dissatisfaction, as the examples did not contain and instructions did not clarify enough to frustrate the annotators.

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?	With only 8 labeled cases of each type, we can introduce a bias in the model where we can be treating exceptions as the general one. In addition, as the x-rays do not follow a pattern, we may have errors due to the image being different from the pattern shown as correct. The results of the initial tests will help to reduce these uncertainties.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Interview a specialist doctor to identify the items checked on the x-ray and compare with the questionnaire questions. Feed the model with more cases labeled with user feedback on cases of false positives, false negatives, true positive and true negative.