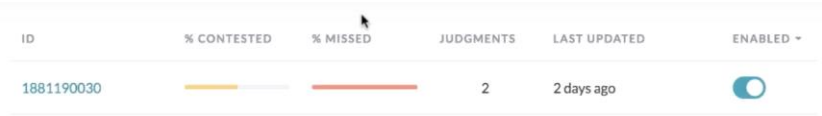
Project Proposal

Boglarka Pankucsi-Szabo

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 the leading cause of death in children worldwide, but in addition to children, the elderly and debilitated people, it can also strike at otherwise healthy young adults. The tenth most common cause of death even today is pneumonia. Inflammation of the lungs (pneumonia) is most often caused by a pathogen, most commonly of viral or bacterial origin. Less commonly, irritating gas, smoke, or allergenic dust, spray, fungus can also cause pneumonia. Our goal is with ML that the doctors, nurses can filter out the people suffering from pneumonia, and the sooner they can start treatment as soon as possible. This can save **lives**, human resources, and money.
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 a binary classification: “yes”, “no”, and “unknown” as options. “Yes” means the presence of pneumonia. “No” is the normal state lungs. “Unknown” is for the cases, where the annotator can’t decide. This can avoid false positives. If “yes” is choosen, it is required to decide how sever is the case. For this I build into the cml a rating field, which show only in the case of “yes”.

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?	10 test questions were made. 7 answer were yes (50% very serious, 38% not too serious, and 13% mild). 2 answer were no, and 1 was unknown. But "it is recommended to have between 50-100 test questions in a job. Because contributors can only see a test question once, the more that are created" *(https://success.appen.com/hc/en-us/articles/213078963-Test-Question-Best-Practices)*
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?	 First I would check if the question is understandable. If not, I would improve it. It could be the annotators difficulties of understanding, which can be avoided by simplifying the language of the question.
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 would give more examples for each label. And reword, clarify the rules and tips.

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 data source should be much bigger. Should contain few thousand pictures. (If we consider this problem as an average problem, then needed 10.000-100.000 images). Images should have the same resolution, size, and zoom.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Check from time to time the questions, and the quality of the answers. If necessary, change, or refine the questions. Rules and tips also might need to be updated given more example pictures.