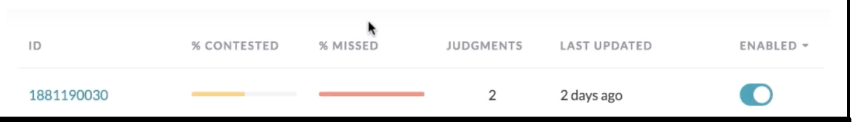
Project Proposal : Medical Image Data Modeling Annotation Project (Soumya Krishnamurthy)

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Disease diagnosis can be revolutionized by machine learning and AI. Images like X-Rays can be effectively classified to accurately distinguish between healthy and diseased organs, like lungs. This has potential to be applied in biomedical imaging and save time and money. This project is used to quickly identify pneumonia in lungs of children. Pneumonia is usually caused by bacterial infection of virus (COVID-9. It shows in X-Rays as areas of cloudiness/opacity in several concentrated areas or one large area. There might also be a general pattern of opacity that obscures the structure of the lungs, heart, and diaphragm. This is caused by swelling (inflammation) of the tissue in one or both lungs. This makes the project a great choice for image classification using ML.
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 use three options: Yes, No and Unknown/Unsure. Since this is an image classification project, the obvious choices are YES/NO. But there might be ambiguity due to quality of images, lack of expertise from the user or unclear instructions/examples. I added a third option Unknown/Unsure.

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 added 8 questions as that was the limit but since there are 118 rows, ideally there should be 10% or 12 questions. Of the 8 questions, 3 were labeled “Yes“; 3 “No” and 2 “Unknown/Unsure”.
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?	 • Include more and better examples • Provide clearer, simple, and relevant instructions • Rephrase questions to be more detailed; have clear answer • Make sure images used are of good quality

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.)	,m  The areas that can be investigated are Ease of job and Fairness of Test Questions. The poor quality of X-Ray images is the first possible issue. Lack of specialist/domain knowledge may mean that annotators are not confident to decide if pneumonia is present. More examples and unambiguous, clear rules in simple language will help take this decision. Fairness of questions can also be improved with better examples and clearer instructions.

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?	From what we now, the dataset is from April 2019, only children's images and from the same hospital. We do not know the gender, age or other demographic information. We do not know how many positive vs negative cases are included. Most importantly, the dataset is limited to 117 rows. This can create many biases in the machine learning algorithm. Enlarging the dataset, good distribution of cases and getting additional information can improve the dataset
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Review questions – from the actual annotator feedback, we can get feedback on prediction accuracy and improve quality of instructions, rules, examples, and questions Review data – Get bigger, good quality data from multiple sources. Improve contributor satisfaction – Iterate based on feedback to get better score.