Project Proposal 

#### *Olena Kormachova*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | The goal is to help doctors quickly identify cases of pneumonia in children from provided images.  With ML support doctors will be able to aid in quickly identifying healthy patients and surfacing potential cases of pneumonia, or reconsider their decision based on the ML suggestion. |
| **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? | There will be three labels:   * Yes – pneumonia symptoms are present * No – pneumonia symptoms not found * Unknown – not sure if symptoms are present or not   As my project is to simply label each image, I have chosen “yes”, “no” labels and added “unknown” choice to capture uncertainty in an annotation.  I could also use a scale of certainty from 0 to 10 to add annotators confidence in their choices, but I decided not to overcomplicate the annotation for this job. |

# 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 developed 8 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? | I would review the rules to make sure they are clear and straightforward. In addition, I would add text field to every answer button where annotators can explain their choice. |
| **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 try to improve Examples area to make sure the annotators have good understanding of the rules.  Additionally, I would ask medical professionals to provide their ideas about improving the annotation job. |

# 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 size of the data which is included in this project is too small and too less of labels for ML model to learn patterns. One of the most common causes of bias in ML algorithms is that the training data is missing samples for underrepresented groups.  To improve the ML model more images should be provided with different sizes, color shades, more examples of present pneumonia symptoms. Like a human learns from the experience of multiple images, the ML model needs to have large pool of examples of labels to improve its predictions. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | With new samples I could improve test questions and examples to show unclear cases. I could add colors to point to exact symptoms in the pictures. |