Project Proposal 

#### *Vibhor Nigam*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | To build a product that can help doctors quickly identify 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? | Decided to mark all areas of cloudiness present in an image and then mark the image as Pneumonia if one or more area of cloudiness are present. |

# 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 a total of 8 test questions. The task is a simple binary classification task and I have included 4 examples from each class in the 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? | Add more instructions on identifying area of cloudiness |
| **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.) | Both examples and test questions. |

# 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? | Given the nature of the data, recall at the cost of precision is expected and thus even minute areas of cloudiness can result into the image being marked as case of pneumonia. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | Add more boundary line images in the dataset where it is difficult to determine with certainty if there is a case of pneumonia. |