Project Proposal 

#### *Wojciech Czuba*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | Goal of this project is to design an annotation job which eventually be used for preparation of a labelled dataset of medical images from chest x-ray, categorized into two classes: healthy and with pneumonia (without distinction on bacterial/viral). Such dataset can be used for developing an ML based classification system which can aid doctors with various experience to quickly identify cases of pneumonia in children.  Due to high complexity of the input data (images - unstructured) it is considered to be not feasible nor do efficient to prepare set of straight rules or develop features manually for a traditional rule-based expert system and so, a deep learning algorithms, such as CNN are preferred solution.  Effectiveness of such system will base on accuracy and quality of the provided data, including appropriate labels, hence proper design and managing of annotation job is critical. |
| **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 are two labels to choose from: 0 - normal (healthy) and 1 – pneumonia (unhealthy) |

# 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? |  |
| **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? | <your text here> |
| **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.) | <your text here> |

# 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? |  |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? |  |