# Project Proposal

**Team member's details:**

- Group Name: IntiLP (Intelligent Language Processing)
- Group Members:
  - Disha Lamba
  - Kiarash Rastegar
  - Somasundaram Palaniappan
  - Mercy
- Email:
  - Disha: jb.dishalamba@gmail.com
  - Kiarash: krastegar0@gmail.com
  - Somu: plsaran97@gmail.com
  - Mercy: creativeart19@icloud.com
- Country:
  - Disha: New York
  - Kiarash: California
  - Somu: England
  - Mercy: London
- College/Company:
  - Disha: MS in Computer Science and Engineering @ New York University
  - Kiarash: MS in Bioinformatics @ San Diego State University
  - Somu: MS in AI and Robotics @ University of Hertfordshire
  - Mercy: MSc in Finance @ UCL
- Specialization:
  - NLP

**Problem description:**

With the ongoing struggle for combating cancer, it has become increasingly important to find underlying causes that may be related to the onset of the disease. Researchers in the biomedical field are constantly looking for new biological markers that may be incorporated into their cancer detection assays. Here we define a biological marker as either a gene or protein that has been observed in that specific disease or cancer. One major issue is that scientists must read through large amounts of articles to find new candidates of biological markers for their cancer assays. Here we propose an approach to use a deep learning NLP model that may be able to extract important information from biomedical literature and return a list of biological markers related to that disease. Here we will focus on 3 major types of cancer: lung, pancreatic, colon.

## Data understanding:

The data we are going to analyze are biomedical literature that can be found on PubMed or other global medical networks. Some problems that we might run into is that not all the articles in the dataset are going to be related to the types of cancers we are interested in. To deal with this we will have to find a specifically tailored dataset related to the cancers that we are interested in. Another problem that we will run into is to remove stop words and clean the text so that we do not have noisy data and our model is not biased towards stop words. To do this we will use Python's Natural Language Toolkit (NLTK) which has functions to remove stop words, normalize text, and remove Unicode characters, etc.

## Project lifecycle along with deadline:

**Week 8: 26 August 20222**

- Loading the data
- Examining the data
- Data EDA analysis

**Week 9: 2 September 2022**

- Data cleaning
- Data manipulation
- Data transformation

**Week 10: 9 September 2022**

- Tokenization
- Padding

**Week 11: 16 September 2022**

- EDA presentation
- Proposing modeling technique

**Week 12: 23 September 2022**

- Model selection
- Building encoder and decoder
- Train, validate, and test the model
- Prediction analysis

**Week 13: 30 September 2022**

- Model iteration
- Experimenting with different hyperparameters
- Final project report and code

**GitHub Repo link:** https://github.com/dldisha/IntiLP