# **Project: Understanding the Determinants of Health**

Team members: Gagan Kaushal, Chengyuan E, Tiam Moradi

Date: 2/13/20

**What is the motivation of the project? (edit for more precise answer )**

BU School of Public Health has access to clinician notes collected during routine medical visits at the Boston Medical Center. These notes are written free-text reports of visit capturing the clinician's findings, observations, and diagnoses.

We need to predict a diagnosis of substance abuse, alcohol use and tobacco smoking from the data in the notes by using techniques such as NLP and machine learning methods.

**What readings did the client suggest for better understanding the problem?**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765598/>

Project background: summarize information from the readings

* Developed a ruled based NLP system that was able to detect substance use statements for alcohol, drugs, and nicotine
  + Also able to obtain semantic elements including:
    - Amount,
    - frequency,
    - type,
    - status,
    - method
    - Temporal (challenges due to variability)
      * There has been limited work on receiving frequency, method, and temporal elements
* There is no baseline for the system this particular group had when doing their study
  + Plan to release a baseline corpora to compare
* Most established clinical NLP systems focus on extracting named entities
  + Diseases, medications, or procedures
* The NLP system performed the worst with Alcohol use statements, performed the best with Drugs
* Substance use statement detection
  + The text was extracted and split into sentences, then the tokens were normalized based on a Lexicon and substance use was extracted.
    - Iterative error analysis on the training dataset improved the system performance.
      * Rules were modified / included to capture cues from section headers/text
  + Results were good
* Elements Extraction
  + Fist patterns were searched and lexical terms that were represented by expressions
  + A set of rules was created to determine if a searched phrase is related to the particular substance of interest by using iterative error analysis on the training dataset.
    - Some elements needed more rules to validate which substance was associated with a phrase
      * 5-15 rules
  + Results could improve

What is the potential approach to addressing this problem?

The project can be split into 2 steps:

1. Use relevant **NLP** techniques to extract information from the **notes** and **form well-structured data** out of it.

(analyzing the notes written in free format as well as tables indicating certain symptoms and patient history.)

1. Build a **machine learning model** to **predict diagnosis** from the data in the **notes**.

What are alternative approaches or techniques given challenges such as, lack of time?

* Visualize a plot of the most frequent medical terminology that appears in the dataset.
* Try standard classification algorithms before attempting deep learning methods
  + Naive Bayes, Random Forests, Support Vector Machine, etc
* Utilize more standard TF-IDF to determine key terms from the clinician’s notes of each patient.
  + We could also try other methods such as word embeddings with doc2vec
  + We could try various preprocessing methods and compare the performance of different methods
    - Lemmatization, Stemming, Tokenization, removing stop words.
* We could use a Recurrent Neural Network using Long Short Term Memory cells on the clinician’s notes and able to take into account sentence structure and possibly take into account what was said previously in the diagnosis.
* We could potentially use topic modeling to cluster patients together; this could help visualize certain diseases be correlated to various key features that were extracted from the clinician notes.
  + We could use Latent Dirichlet Allocation for this.
  + Could compare to KMeans++ for results.

**What are the expected outcomes? What is the most pressing question? What does the client define as “done”?**

The expected outcome is an estimate of the predictive accuracy of the method.

**What are the open questions or uncertainties about the project?**

* How big is the dataset?
  + How many years worth of notes are we going to be training our model on?
  + Are these notes all from the same hospital or from various hospitals?
    - Different professionals may word the clinician’s notes differently.
* How many different diseases are we expected to be able to predict from in the dataset?
  + Is this dataset balanced in this regard?
  + Is the dataset labeled?
  + What is the level of detail for the classification (ex: cancer )
* Will there be any additional dataset/resource available to us to better understand the complex medical terms mentioned in the dataset?
* Does patient history include a family history of chronic diseases?
  + What format is the data?
  + What is the type of data? (Strings, Quantitative, etc )
* Have you researched this problem objective?
  + If so, what were the results of the previous research?