The overall vision of HLP is to use health information from various sources to derive insights for improving public health. In this theme, our broad objective for this year’s internship program is as follows. which we plan to explore as a team in the coming weeks.

**Objective: Develop and evaluate pipelines using multi-task learning**

Increasing adoption of technology across countries and age groups have resulted in widespread availability of information. Such information has been vital in studying common patterns for expanding our knowledge and extracting important signals for drawing meaningful conclusions. Data sources like Twitter adds more than a million tweets everyday which presents interesting challenges. Firstly, the amount to data is too large to be processed manually and hence require automated methods to extract interesting signal from noise. And secondly, data spans across a wide variety of topics thus making it challenging to identify the tweets of interest which require automation as well. Machine learning provides a family of tools used for identifying interesting patterns in large amounts of data.

Increasing focus of machine learning has resulted in parallel research efforts across teams resulting in multiple annotated datasets with similar annotation guidelines for building machine learning models. In traditional machine learning we train the model on just the data annotated for the task, however multi-task learning, we can use data from related tasks in addition to the specific task to nudge the models to do better. This provides us the opportunity to train a common model using multiple datasets such that the model can learn from all datasets collectively and generalize across tasks. We will experiment with this idea across two different pipelines for which we will build applications. We will treat them as individual case studies as follows:

**Case Study 1: Detect personal declarations of COVID cases.**

With increasing attention on COVID19 monitoring on social media, many datasets are available that find personal mentions of the disease. Most of these datasets have similar annotations, and hence we can use such annotations for multi-task multi-corpus training to develop better models.

**Case Study 2: Detect expressions of diagnosed health conditions.**

We currently possess annotations from multiple datasets for extracting health, medication, symptoms, adverse drug reaction (ADR), indication and drug intake related tweets across independent datasets. The goal here is to develop a single multi-task system for extracting all signals by collectively training all datasets.

As part of Case Study 2, we will also be participating in 2 shared tasks as part of the [WNUT-2020](http://noisy-text.github.io/2020/) workshop at EMNLP Conference: <http://noisy-text.github.io/2020/>

1. Identification of informative COVID-19 English Tweets
   * Task description : <http://noisy-text.github.io/2020/covid19tweet-task.html>
   * This is a classification task where each tweet is tagged as informative and uninformative
2. Extracting COVID-19 Events from Twitter
   * Task description : <http://noisy-text.github.io/2020/extract_covid19_event-shared_task.html>
   * This is a slot-filling task with a prior baseline which we will seek to improve

***Skills:***

*Fulfilling both objectives will involve various aspects of data annotation, data analysis, classifier construction for prediction, pipeline development and UI design. You will be expected to perform one or more of these tasks as part of the internship based on your interests and current skills.*