Problem Statement

Title	**Patient Case Similarity**
Description	The objective of patient case similarity is to identify similar patients based on their EHR data. Identification of similar patients cases be useful for improving patient outcome, for treatment or drug recommendation to a new patient, prediction of clinical outcome, clinical decision support, research on those cases. **Task**: Applying machine learning algorithms to find similar patient cases from given dataset. This dataset contains a mix of structured and unstructured data from patients' EHR.
Problem Code	**RG2**
YouTube Link	**https://www.youtube.com/watch?v=skDcDRBdfNI**
Nature/Difficulty	Complicated
Category:	Software - Healthcare
Technology Bucket:	Software – Natural Language Processing, Machine Learning, Deep Learning
Sample Data	Patient Demographic CSV Text Files NER XML Files (for above text files)
Actual Data	Will be provided in Hackathon
For more details, contact	Vivek Kumar (**vivek.k@ezdi.us**)

Detailed Description

Introduction:
An electronic health record (EHR) is the systematized collection of patient health information in a digital format. To effectively use the increasing Electronic health record (EHR) data, patient case similarity is very important. Structured and unstructured data of patient cases can be used find similar patient cases. The objective of patient case similarity is to identify similar patients based on their EHR data. This can be used for improving patient outcome, for treatment or drug recommendation to a new patient, prediction of clinical outcome, clinical decision support etc.

Other keywords: Case-Based Reasoning

Tasks:
Aim of the problem is divided in two parts:-

1. Predict similarity score for a given patient for all other patients.
2. Find factors or evidences that indicates or supports the above similarity scores.

Dataset Description:

- Patient Demographics and Case Details (.CSV File)
    - This file contains all demographics information about patients
    - Demographics information may consist of features like patient id, patients Age, patients Length Of Stay, patients Locality,...etc.

Sample CSV:
https://docs.google.com/spreadsheets/d/1ke3F7aJdLqgG110lJKVte8X91peTs1vBM7eTMo3e5zw/edit#gid=0

- Raw text files
    - This folder will consist of text files of patients dictated by the physicians over the course of the case.
    - One patient may have more that one file.

Sample Raw Texts
https://drive.google.com/open?id=1xZVkVKSg3UJuFdQ5qfgeaMUew50Hl48f

- NER output (**.xml** format)
    - For each raw text file, an xml file containing the recognized Named Entities is present.
    - Description of xml format:
        - Each file is divided into various **Section** tags
        - Every section can have multiple **Paragraph** which in turn can have multiple **Sentence** which will contain the detected entities like:
            - relationSet
            - problem
            - anatomicalStructure
            - lab
            - modifier
            - medicalDevice
            - finding, etc
            - Details about the above annotations can be found here: https://docs.google.com/document/d/10U8QaT7wq5iRjiToW88da4T9SrfcaC3d69fPBbiOu34/ under "**Description of Annotations".**
        - **begin**, **end :** Beginning and ending index of the entity, calculated from the start of the file
            - If a detected entity is split over different words, then the begin and end will have multiple entries separated by "**_SEP_**"
        - **status:** Temporal status of the entity.
            - 0 = The patient presently has that
            - 1 = Detected as past history of patient
            - 2 = As family history of patient (not his own)
        - **certainty**: Negation certainty of detected entity
            - -2 = Negative
            - -1 = Possibly negative (Not sure)
            - 0 = Affirmative (Confirmed positive)
            - 1 = Possibly positive
        - **cui, tui, rank, sourcetype:** Ordered results containing the identifiers of the entity from the UMLS Metathesarus.
            - **UMLS :** https://www.nlm.nih.gov/research/umls/

Sample XMLs (Corresponding to the text files above**)**
https://drive.google.com/open?id=15zx8tq_i2qqdLctW5GlK0rshlKWCZIMc

Supporting Material:

Research Papers containing similar problems:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5121278/
- https://medinform.jmir.org/2017/1/e7/pdf
- https://www.computer.org/csdl/proceedings/bibmw/2010/8303/00/05703846.pdf
- https://par.nsf.gov/servlets/purl/10026424
- https://arxiv.org/pdf/1704.07498.pdf

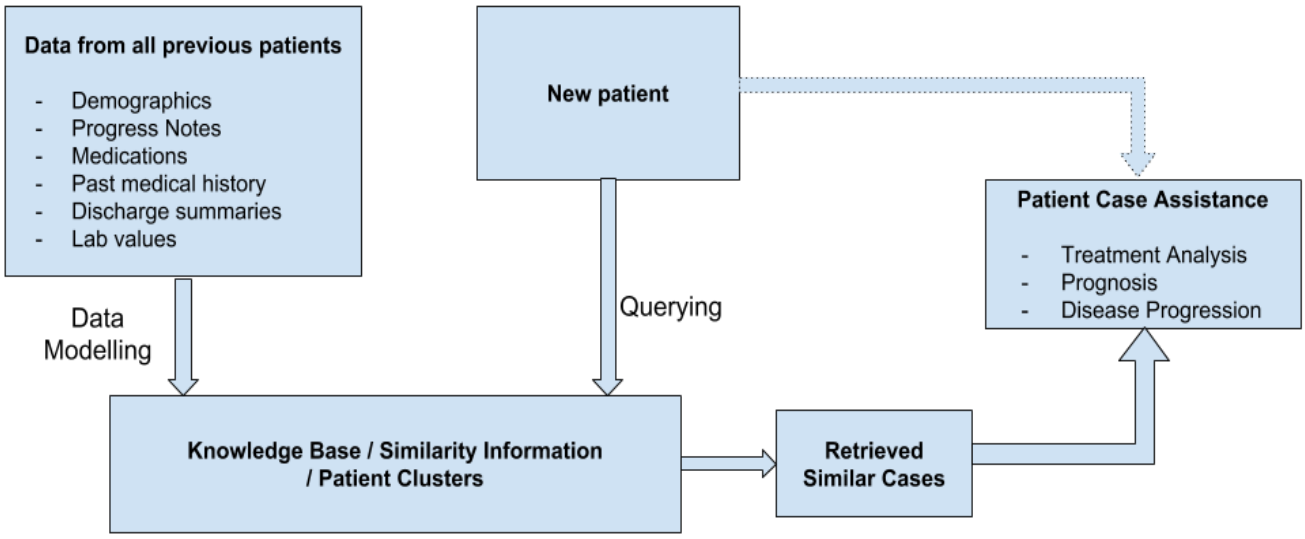
Medical Datasets:

- **MIMIC-III** (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units

  Link: https://mimic.physionet.org/about/mimic/

Frequently Asked Questions:

1. **Describe more about this problem statement as a machine learning problem?**
   Answer:
   The intuition behind patient case similarity is that if two patients are similar (based on a number of facts), then their medical case progression should also be similar. Identifying past patients which are similar to the current patient under consideration and deriving insights from them to investigate diseases and potential treatments. As such this becomes an **unsupervised learning problem** where the main goal is to derive a similarity measure between the patient cases and then **cluster** those cases according to their similarity / distance.



   For example, one can use the given demographic data along with the vector representation of textual data to generate a numerical representation of a patient. This representation can be tweaked by giving different weights to different entities like what diagnoses the patient had, what procedures were performed on him, what medications he was taking etc. To help with those, XML files containing the NER entities are also supplied in addition of text data.
   This model (which can take the given inputs for a patient and then convert it into a vector and then apply a similarity measure between pair of patients to calculate the similarity between them) can be stored as a knowledge base. So when a new patient comes, we send it to the knowledge base, which then calculates the similarity of new case with all older cases and give out a few most similar cases.

2. **What type of data is given? Where can I find similar data? What is MIMIC? How is it helpful here?**
   Answer:
   The given data is a part of EHR data of a patient case. When a patient comes to a hospital with some problem, the hospital is required to keep a record of the progress. This is done in EHR. More information can be found here:
   - https://en.wikipedia.org/wiki/Electronic_health_record
   - https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html

   This data contains the observations made by the physician / nurses about the condition of patient in different document types. For example:
   - An **Admit Note** may contain the first communication between the patient and physician and the chief complaint with which the patient presented to the hospital.
   - A **Progress Note** will contain the daily / routine checks on the condition of patient, along with lab measurements / tests on the patient.
   - A **Procedure / Operative Note** will contain details about a surgery performed on the patient.
   - Finally **Discharge Note** is made once a patient is discharged from the hospital. It may contain a brief summary about all the things that were observed, and carried out on the patient during his case.

   Note that not all cases will contain all the above mentioned type of notes.

   Since medical data of patients is very **sensitive and confidential**, it needs to be **protected**. And so, there are only a few open datasets for medical data and most require an extensive certification and licenses along with extensive **de-identification** of datasets to remove the patient's information which is very time consuming.
   See this link for more information:-
   - https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected

   MIMIC-3 is a similar dataset which is freely available after you do a course successfully and apply for access. You can use MIMIC-3 dataset or any other similar medical dataset to:
   - pre-train the models
   - training the word-embeddings,
   - relationships between diagnoses of patients,
   - reviewing / implementing related research papers
   - any other task which may help in patient similarity

   But **keep in mind** that the evaluation of your solutions will be done on the based on **our dataset** which will be provided during the hackathon.

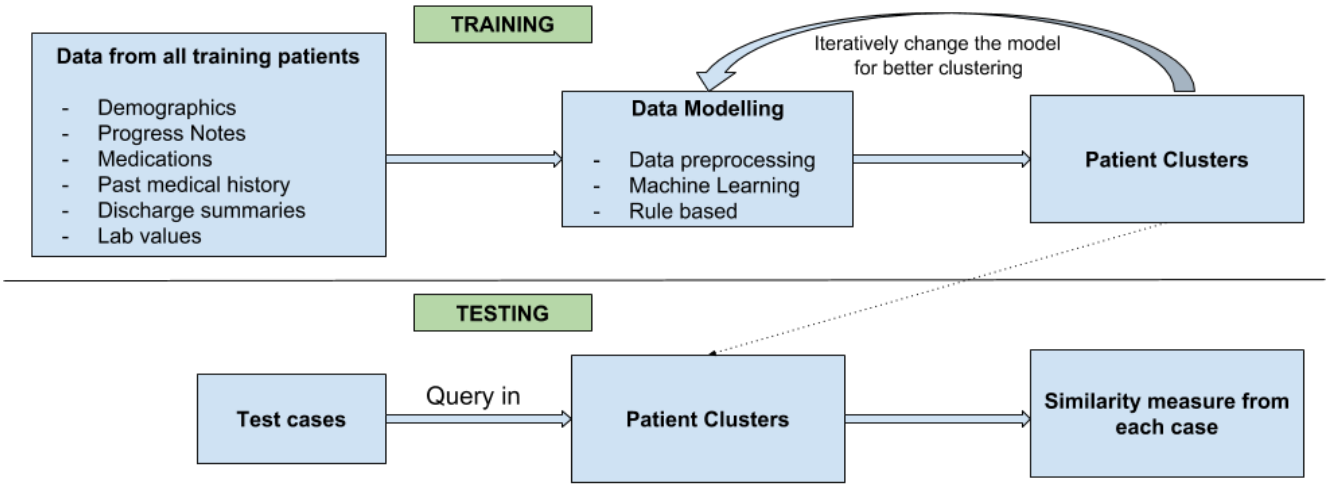
3. **How is the sample data helpful? Is it enough for the problem?**
   Answer:
   Sample data has been uploaded in the documentation to give you a sense of the actual data and what to expect in it, so that you may prepare the necessary **preprocessing code** like **cleaning, xml parsing**, and see what **type of entities** are in it. Since this is a ML based problem, the actual dataset and model training should be done in the hackathon only.

4. **What will the actual input format? Why are three types of files present in sample data?**
   Answer:
   The format given in sample data provided is same as actual data. You will get a single csv file containing the demographic data of the patient along with the an **id.** This **id** can be used to select the text files for a given patient from all the text files given. The xml files will use the same name as text files to indicate their mapping.

   You may choose to use demographics + text or demographics + xml or all three in your unsupervised algorithm.

5. **When will actual data be given? Is there any marks for UI development (app or website) for displaying the output?**
   Answer:
   The actual dataset will only be provided in the hackathon to the shortlisted teams. The reasons are:
   - This is an ML based problem and there are no marks for the app or website development. All marks are for the solution only
   - As described in Ques-2, the dataset is sensitive and needs to be protected.

6. **Will the actual data include training and testing data both?**
   Answer:
   Only the training data will be given to the participants. You can perform experiments on that to make your model. We will take your scripts (model + processing) and run it on test cases to get the score.



7. **What is the required output format?**
   Answer:
   You need to output a similarity matrix between the given cases. For example if we give 100 cases in training (C1.. C100) and 20 in testing (CT1..CT20). So you will give two matrices as output:-
   - **Similarity matrix between the 100 cases**.

	C1	**C2**	**...**	**C100**
C1	1	0.8	...	0.4
C2	0.8	1	...	0.35
...	...	...	...	...
C100	0.4	0.35	...	1

   - **Similarity matrix between the 100 training cases and 20 test cases**.

	C1	**C2**	**...**	**C100**
CT1	0.4	0.8	...	0.45
CT2	0.8	0.68	...	0.65
...	...	...	...	...
CT20	0.4	0.35	...	0.15

8. **What is our expectation from the proposal?**
   Answer:
   We will be looking at the **feasibility** and **practicality** of your decided approach.