PROJECT PROPOSAL REQUIREMENTS
1) Identify and motivate the problems that you want to address in your project.
2) Conduct literature search to understand the state of arts and the gap for solving the problem.
3) Formulate the data science problem in details (e.g., classification vs. predictive modeling vs. clustering problem). 4) Identify clearly the success metric that you would like to use (e.g., AUC, accuracy, recall, speedup in running time).
5) Setup the analytic infrastructure for your project (including both hardware and software environment, e.g., Azure or local clusters with Python, PyTorch and all necessary packages).
6) Discover the key data that will be used in your project and make sure an efficient path for obtaining the dataset. This is a crucial step and can be quite time-consuming, so do it on the first day and never stops until the project completion.
7) Generate initial statistics over the raw data to make sure the data quality is good enough and the key assumption about the data are met.
8) Identify the high-level technical approaches for the project (e.g., what algorithms to use or pipelines to use). 9) Prepare a timeline and milestones of deliverables for the entire project.
10) It's required to utilize deep learning methods say CNN, RNN, GNN etc. in your project.

Notes:
Clinical notes suffer from "curse of dimensionality" Clinical notes also exhibit a hierarchical sequential structure: a longitudinal patient record includes a time series of notes, each itself consisting of a sequence of words. Framing the problem as a temporal problem
e.g. Using Patient A clinical notes sequentially to predict progression of disease, co-morbidity etc

Can be used for annotation:The National Center for Biomedical Ontology (NCBO) Annotator (LePendu et al., 2013), which extracts occurrences of terms in an expansive vocabulary of biomedical terms compiled from a collection of controlled terminologies and biomedical ontologies.
Can be used for normalization unique biomedical concepts using the Unified Medical Language System MetaThesaurus, which provides a mapping of strings to Concept Unique Identifiers (CUIs)

Problem Statements
Compare different embedding techniques including word2vec(which are word level embeddings) and comparing it with say document level embeddings(patient level embeddings), understanding how different embeddings can affect prediction models.
Papers
file:///Users/boshikatara/Downloads/Learning_Effective_Representations_from_Clinical_N.pdf
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2744372.pdf

modeling high-cost tasks like ER, ICU stay durations using in-hospital mortality and stay duration using clincial notes. We can also frame this problem as co-morbidity problem using MTL

https://www.nature.com/articles/s41467-021-20910-4

http://tjn.mit.edu/pdf/whats-in-a-note.pdf
MTL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568068/
https://psb.stanford.edu/psb-online/proceedings/psb19/ding.pdf