**Idea 1:** <https://www.kaggle.com/piotrgrabo/breastcancerproteomes>

**Description of dataset and problem statement:** This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). It contains expression values for ~12.000 proteins for each sample, with missing values present when a given protein could not be quantified in a given sample.

Developing a machine learning model to detect proteomes responsible for breast cancer and its subtypes can help in identifying and categorizing cancerous and non-cancerous patients.

**Idea 2**: <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

**Description of dataset and problem statement:** This dataset will be used to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This allows the worldwide AI research community the opportunity to apply text and data mining approaches to find answers to questions within, and connect insights across, this content in support of the ongoing COVID-19 response efforts worldwide. There is a growing urgency for these approaches because of the rapid increase in coronavirus literature, making it difficult for the medical community to keep up. The target is to develop text mining tools.

**Idea 3**: <https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset>

**Description of dataset and problem statement:** Fake news spread misinformation, disinformation or mal-information. They spread through word of mouth, traditional media, digital forms of communication such as edited videos, memes, unverified advertisements and social media propagated rumors. They have potential of resulting in mob violence, suicides, people following misinformation during pandemics, panic spreading, stampedes etc. I came across a dataset at Kaggle which can be used to develop models based on natural language processing to detect fake news. Developing models to develop approaches to detect fake news can help in detecting what is fake and true. It can save a lot of financial hazards, social and miscellaneous havocs. This dataset contains 40000 articles consisting of fake as well as real news. The fake and real news data is given in two separate datasets with each dataset consisting around 20000 articles each. The target is to create and train a machine learning model so that it can correctly predict whether a given piece of news is Real or Fake.