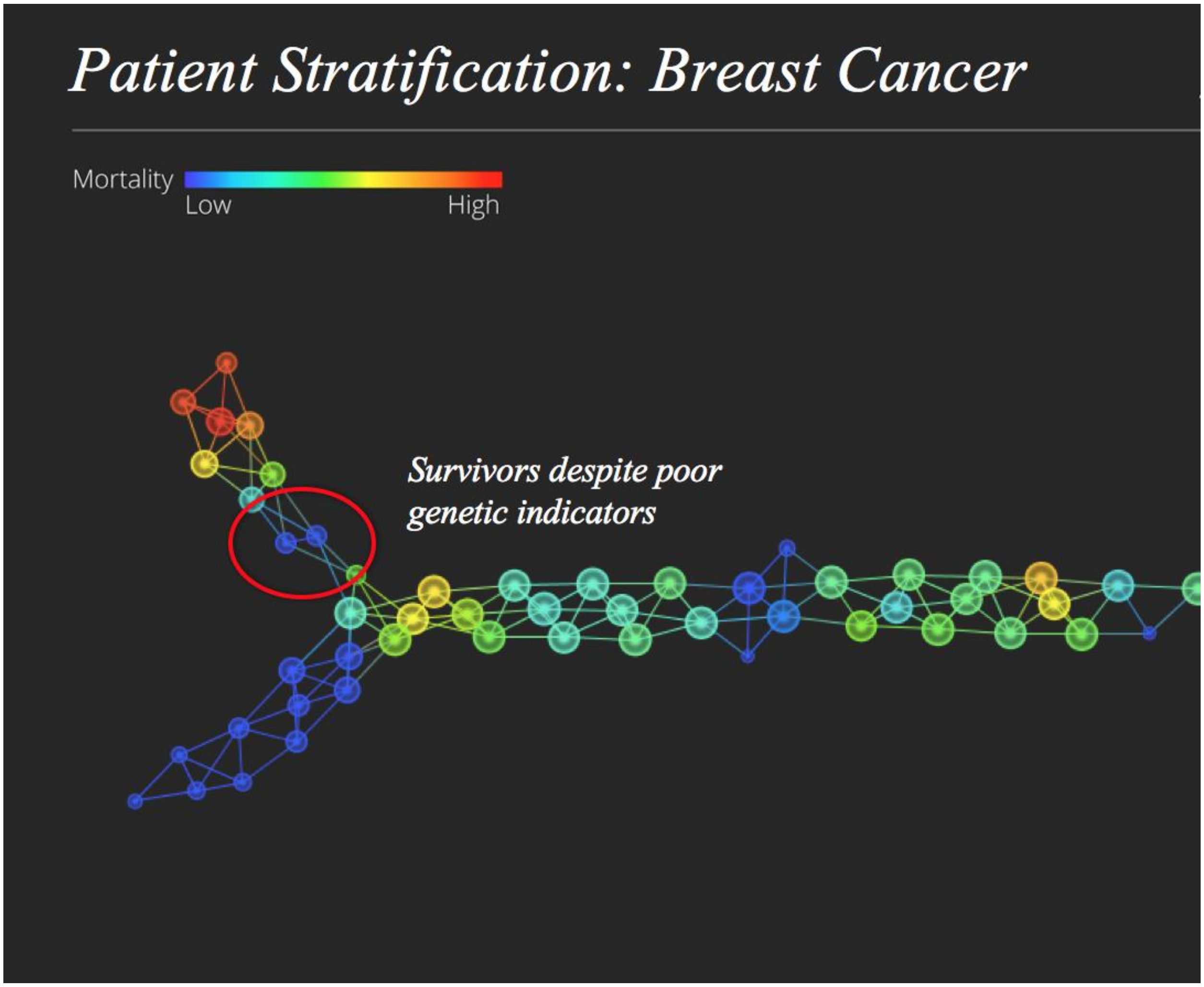
NKI Breast Cancer Dataset Summary

NKI Breast Cancer dataset contains data for 272 breast cancer patients with 1570 features. You can build the following network shown in the below figure using the gene expression values given in the dataset. Some metadata includes patient info, treatment, and survival.


Patient Stratification: Breast Cancer
Mortality
Low
High
Survivors despite poor genetic indicators

Each node is a group of patients like each other. Flares (left) represent sub-populations distinct from the larger population. (One differentiating factor between the two flares is estrogen expression (low = top flare, high = bottom flare)). Bottom flare is a group of patients with 100% survival. Top flare shows a range of survival–very poor towards the tip (red), and very good near the base (circled).

The circled group of good survivors have genetic indicators of poor survivors (I.e. low ESR1 levels, which is typically the prognostic indicator of poor outcomes in breast cancer)—understanding this group could be critical for helping improve mortality rates for this disease.

You could find the data here. More details about the features in the data are available in the research paper: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival

Using the dataset named **NKI_cleaned.csv** and resources available (a helper notebook is also available in the link here for a quick start), attempt the following exercises

1. In the following two exercises, we will look for your ability to extract details from various sources of information and break down complex problems

    a. Write a brief note on the dataset and the research work from where this dataset originates. We expect no exploratory analysis.

    b. How do you think the creators of this dataset have calculated the values in column 17 to 1554? And what does the column names mean? Write a brief explanation based on your understanding and study.

2. In the following two exercises, we will look for your ability to work with data and build machine learning models

    a. Perform some exploratory analysis using the dataset which helps in understanding the data and the problem at hand. The notebook in the link gives some starting point.

    b. Choose 3 of your favorite machine learning models apart from the one given in the notebook and apply it on the dataset. Perform an analysis to compare the models using the model performance metrics. Also, show if hyper-parameter tuning improves the results in your comparison. You are free to build an ensemble model if you think it is going to be better than the individual models.

3. In taking a data science project from prototype to production, highlight at least five challenges that you have faced in your past projects.

General Instructions:

1. Use Google Colab notebook to submit your solutions (even for the first two and the last exercise).
2. We strictly prohibit plagiarism. Any copy paste work will disqualify you from the selection process.
3. There will be no guidance provided for the exercises. Wherever there is some ambiguity, make suitable assumptions.