Project Outline:

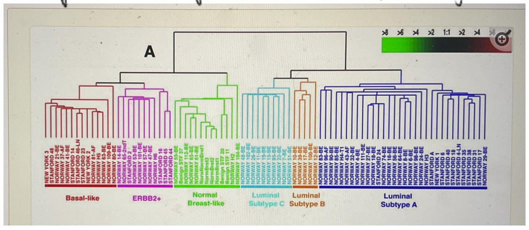
Goal: relate differential mRNA expression in cancerous tissue relative to normal tissue to cancer mortality and recurrance rates.

Pre-processing / EDA:

Ari → 1. Aggregate all patient data based on ID and mRNA Z-scores compared to normal tissue data.

- data_Clinical-patient.txt and data_mrna_seq_v2_rsem_zscores_ref_normal_sample.txt

2. Clustering to determine / confirm established breast cancer subtypes
   - Want to identify clusters of genes which are expressed together



from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC58566/

3. PCA, initial plotting
   - plot overall patient mortality & cancer recurrance
   - plotting dataset histogram by clinical breast cancer subtype & cancer stage
   - PCA & Skree plot of mRNA transcripts

Goal: Done by 4/6

Model Development:

and / or cancer recurrance

End goal: logistic regression to predict mortality based on differential mRNA expression

- Feature reduction: A. PCA and then LASSO
  B. Via forward-stepwise subset selection
  C. LASSO on original dataset

Planning Presentation: (16 min presentation, 8 min Q&A)

- 5-6 min: background & dataset explanation
- 4-5 min: on dataset pre-processing, exploration & model development
- 5-6 min: on model results, interpretation & further discussion of model performance
- 1-2 min: Conclusion