**Capstone 2 Project Ideas:**

1. **What transcriptomic, proteomic, genomic features predict response to a therapy?**

**Why**: Identified features/regressors will work to predict either an enrolement biomarker or a causative biomarker important for drug targeting.

**How/Data:**

Dataset1: Each column indicates a cell sample (cell line/tumor tissue) and each row indicates a feature. For example, Transcriptomic Features ~ 55,000 rows, 100 columns.

Dataset2: Response of a drug for each cell sample (cell line/tumor tissue)

For a single drug; we have to create a response (X= target) versus feature (Y= regressors).

**Ref**: <https://linkinghub.elsevier.com/retrieve/pii/S0092-8674(18)30308-8> (The data is very clean and raw data has already been analyzed. I do have csv format raw data; when opened in excel 54,589X100 table for transcript data)



**Potential pitfalls:**

* N/P ratio problem. Elastic net or regularized regression should work the best.
* Want to test other regression approaches: logistic, RF etc.

1. **Is there a common transcriptomic signature for lung cancer (type 1-4), melanoma types and glioblastoma tumors?**

**Why**: Identified signature would allow for early detection.

**How/Data:** Transcript data is publicly available (NCBI GEO datasets, Depmap, Dataset 1 in project idea 1).

Each column indicates a cell sample (cell line/tumor tissue) and each row indicates a feature. For example, Transcriptomic Features ~ 20,000 rows, 150 columns.

Steps of data analysis required: 1) Merge multiple datasets to create a universal set.

2) Generate heatmap/PCA (SOM) for each samples. 3) Image comparison algorithms to be deployed.

3) 80-20 split. Training dataset= heatmaps with curated melanoma, lung, glioblastoma heatmaps, Test= blinded samples for which labels are available

4) Measure accuracy test

5) To test for ground truth: possible blinded unlabeled samples: would be hard to find

1. Recommending movies and predicting user personality types from Netflix movie (item) vs user clusters using collaborative filtering:

**Why:** This is useful for marketing purposes for customized/targeted advertisements. Furthermore, basic personality information may be useful for basic research.

Dataset: Netflix ratings dataset is available in the form of a curated csv.



User based collaborative filtering for recommendations using cosine similarity and KNN.

Improve the algorithm using hyperparameters search.