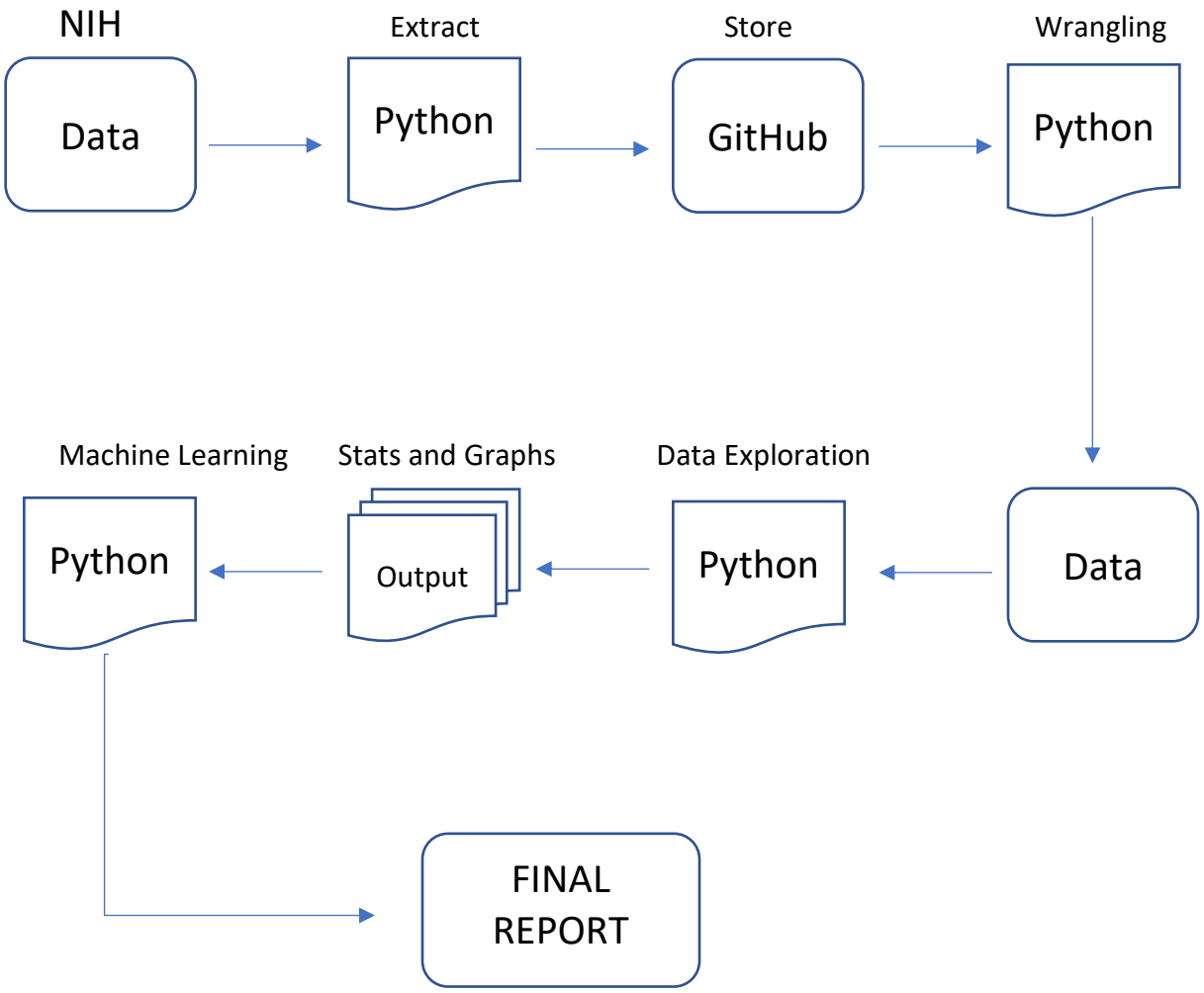
Architecture of Bladder Cancer Project
Flow Chart



Cleaning and Wrangling (Pandas and Numpy)

- remove information not related to analysis
- transpose the data (pd.df)
- Indexing and naming

```
df.loc[0, 0] = 'Gender'
df.loc[0, 1] = 'Age'
df.loc[0, 2] = 'Class'    #this is the class label
df.loc[0, 3] = 'Path_Stage'
df.loc[0, 4] = 'Path_Grade'
df.loc[0, 19] = 'Sample_City'
df.loc[0, 25] = 'Sample_ID'
```

- Out of the 972 serum samples, 392 are labelled as "bladder cancer", 100 are labelled as "non-cancer control" and 480 samples are labelled as "other" types of tumors. We will combine non-cancer and other and label it "Control"

- Each sample has six variables to describe the patient (disease status, gender, age...) plus 2,565 gene measurements. The gene measurement in our downloaded file is log2 converted. We will try to re-store each data point to its original measurements (to the 2th power of the readings), and explore different data normalization methods.

- Normalization method will be used:
    - Assume every sample started with the same amount of genes (sum equal);
    - Assume every sample has the same middle value;
    - Assume same genes constant in every sample;

- Normalization will designate a low value to every reading that is below that value (assume machine detection has a sensible range).

Data Exploration (Pandas, Matplotlib, Seaborn, Statsmodels)

- Summary of the characteristics
- Correlation
- Hierarchy clustering
- Principle component analysis
- K-mean clustering
- Plots and graphs

Machine Learning (SciKit-Learn)

- Split data into Test (35%) and Training (65%);
- Try different algorithms
- Identify top features that contribute to the difference between bladder cancer and other control samples.