**Capstone Proposal**

1. Problem Definition
   1. The objective is to determine the most energy efficient design layout for residential buildings, given a number building parameters and looking at the output heating and cooling load requirements which would directly correlate with the energy efficiency.
2. Significance to Client
   1. Architecture and Civil Engineering firms would be able to use these conclusions to design more optimally structured buildings geared toward energy efficiency. Government departments and organizations learning towards green initiatives would be able to use these conclusions for setting standard specifications and certifications.
3. Data
   1. The data used will be provided by a dataset created by UCI researchers. Twelve different building designs were simulated on Ecotect, an environmental analysis tool for building performance. There are 768 data points with 8 input variables and 2 outputs. This includes relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution, heating load and cooling load. (Will have to defend accuracy/usefulness of Ecotect simulation; alternative datasets with varying inputs also available)
4. Solution Outline
   1. Sort data into useful figures
      1. Ex. Sort from least to most efficient loads
   2. Determine any correlations, trends
   3. Determine strength of input variables as they relate to output variables, linear regression
   4. Use visual data points to present insights on an optimal design; craft a data story
      1. Ex. Include multiple 3D visuals to graphically narrow down optimal parameters
   5. Machine learning - given a new set of parameters, can we predict the heating/cooling load
      1. training set, validation, test set
      2. Evaluate multiple algorithm
      3. Determine which algorithm models the best relationship between variables to predict outputs
      4. What accuracy can we predict
5. Deliverables
   1. The dataset provided by UCI
   2. IPython notebook so that tests can be repeated
   3. Code on github
      1. Used to clean data
      2. Used apply statistical analysis
      3. Used to produce visuals
   4. Visuals relevant to data story and conclusion
   5. Paper write up including description of statistical analysis and conclusive visual data sets