Technical Design

# Data Input

Assumption: Microsoft Excel can handle the volume of data required for this analysis.

The botanist can input the raw data into an Excel spreadsheet, which will be converted to CSV format.

How do we handle missing data and outliers?

# Processes

## Visual Analysis

Use a 3D scatterplot to understand the data and correlation between features

Represent the data in a correlation matrix to statistically determine the relationship between features

## Data Transformation

Use Principal Component Analysis to reduce the number of dimensions; to identify important variables and filter out irrelevant ‘noise’ features from the dataset (note: in the business analysis we only mention 4 variables, in that case PCA is probably not necessary, but I remember him saying a correlation matrix is a hint for this, so maybe there would be more than 4 variables in the dataset? So we could apply PCA in that case? I really don’t know what I’m talking about)

Fit a classifier to the PCA transformed data

## Build a Classification Model

Assumption: we are not provided with the flower type classes -> we use an unsupervised learning algorithm like K Nearest Neighbour to group the flowers based on their features. Graphically represent the data for each k value tested.

Use 70% of the data to train the model, 30% to validate the model, then use cross validation.

# Tables/objects

# Outputs

Flowers are

<https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9>

<http://setosa.io/ev/principal-component-analysis/>

<http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html>

<https://medium.com/@haydar_ai/learning-data-science-day-10-classification-k-nearest-neighbors-and-cross-validation-d7d58dbe1fed>