# In Class Activity

## Semi-Supervised Learning – Roadmap Activity 2

This activity will extend the first Roadmap activity. It will add Dimension Reduction, in the data cleaning step, using a correlation matrix technique.

The result will be a more focused, informative result from the cluster analysis.

Some important definitions:

**Dimension Reduction** – Performed at either the "data cleaning" step or in the evaluation of a model's output, dimension reduction is the act of reducing the number of variables. This should be done to isolate the variable (i.e. dimensions) of interest.

Dimension reduction is sometimes done to make the model more accurate.

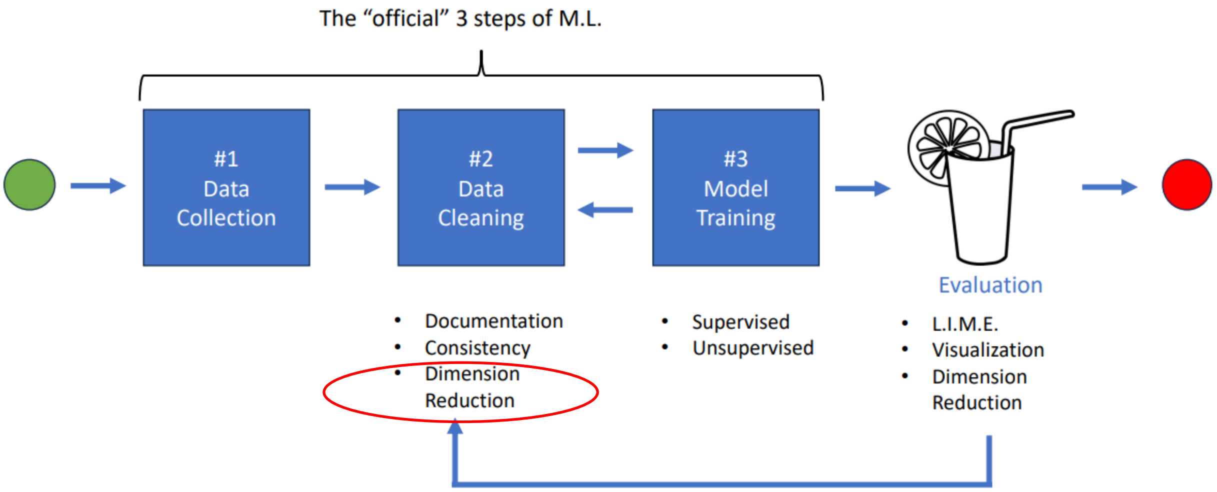
It is also done to make the model more easily understood.

For example, it is difficult for humans to interpret 3 dimensional diagrams, and higher dimensional information is impossible for humans to interpret visually.

But, if the two most relevant dimensions can be identified, then multiple visualizations (scatter plots, histograms, line graphs) are possible.

**Correlation Matrix** – A correlation matrix is a statistical technique. It is not an M.L. model in the same way that Cluster, Decision tree, or Linear Regression is.

# "Shafer's Roadmap"
## *Semi-Supervised* Machine Learning



## Instructions

1. Students should visit http://tinyurl.com/shaferaicourse and download the files found in the **roadmap2** folder.

2. Start by examining flipped-classroom-survey.xlsx and flipped-classroom-survey.csv in Excel. They have already been prepared for you. ***That is, the data cleaning has already been done.***

3. Together in class we will (re)run the cluster script, decisiontree script, and regression script.

4. Now run the correlation-matrix script. Examine the correlation matrix.

   a. A correlation matrix may be generated using repeated ***Pearson*** tests. Pearson tests are used to compare ***continuous*** variables. They assume that the variables being evaluated follow a ***normal distribution***.

   b. A correlation matrix may also be generated using repeated ***Spearman*** tests. Spearman tests are used to compare categorical variables. There is no assumption of a normal distribution.

   c. Which one do you think we are using here? Why?

   d. Using output from the correlation matix script we will identify the two most interesting questions for our cluster analysis.

   e. Edit the cluster script to evaluate only the two most interesting dimensions, and plot the results.

5. To document your participation in this activity, your instructor will ask you to post one of the graphics you generated, along with some your comments regarding it. Post these to the corresponding discussion post on Canvas.

   Be sure to post the graphic that your instructor requires.