# A2 Research

## Semi-Supervised Learning

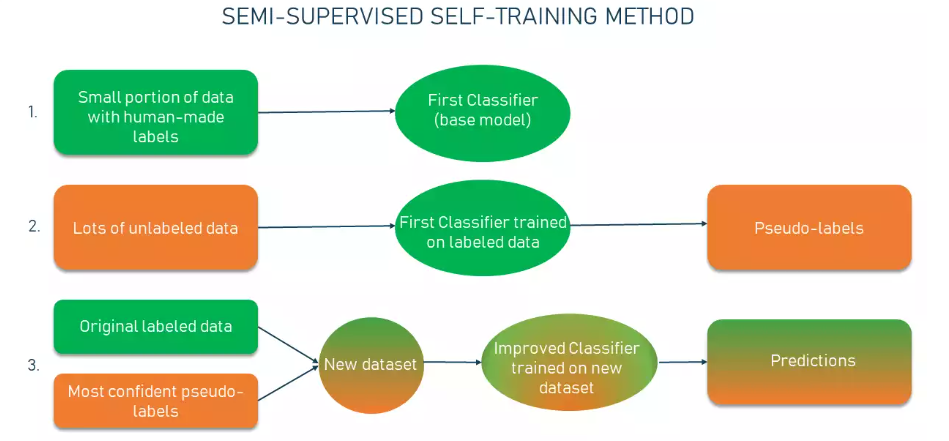
Also called Weak Learning

<https://www.altexsoft.com/blog/semi-supervised-learning/>

Technique to train using both labelled data and also unlabelled data.

The basic premise is this:

1. Train a base model on the labelled data
2. Apply **pseudo-labelling:** 
   1. This is where you predict on some or all of the unlabelled data.
   2. Then, you try to rate the confidence of your predictions.
   3. Find the most confident of the predictions, for example greater than 80% confidence
   4. Take these most confident predictions and combine them with the original labelled dataset.
   5. Then, train a second model, based on both the labelled and most confident predicted data
   6. This process can iterate (10 rounds is common)



There is another similar process called **Co-training.** Haven’t read yet into the details but apparently it’s about training two individual classifiers based on two views of the data

# Research Findings and Report Guide

* Evaluation Metrics
  + IsCancerous, use F1 Score
  + CellType, use Accuracy
* Basic Decision Tree model was generated in file 04, but results were not good. Best not to consider as a Base line Model
* PyTorch Baseline Model 2 Layer Fully Connected Neural Network in File 05c.
  + No Image Preprocessing to grey scale has been done
  + Results: IsCancerous
    - Training Accuracy: 0.961
    - Training F1 Score 0.945
    - Test Accuracy: 0.873
    - Test F1 Score: 0.894
  + Results: CellType
    - Training Accuracy: 0.867
    - Test Accuracy: 0.79
* Tensorflow Base model was created, in Assignment2 notebook.
  + Image Preprocessing applied, converted to greyscale
  + 2 Layers
  + Results: IsCancerous
    - Training Accuracy 89%
    - Test Accuracy: 82%
  + Results: CellType
    - Training Accuracy: 76%
    - Test Accuracy: 70%
* PyTorch 3 Layer Fully Connected Neural Network in File 06.
  + No Image Preprocessing to grey scale has been done
  + Results: IsCancerous
    - Training Accuracy: 0.955
    - Training F1 Score 0.936
    - Test Accuracy: 0.863
    - Test F1 Score: 0.884
  + Results: CellType
    - Training Accuracy: 0.883
    - Test Accuracy: 0.784
* Experimented with More layers in files 07 and 08
* Tuned Learning Rate in file 09
  + Picked 0.0001 as best learning rate
  + 0.00003 has slightly better results in experiments, but was very slow, not worth the time tradeoff