BTP Weekly Report 1
February 2, 2018

---

1 Problem Statement

We consider a scenario where we imagine ourselves as Botanists and we want to classify **Iris** flowers. We chose machine learning as our tool to do the task. For instance, a sophisticated machine learning program could classify flowers based on photographs. Our ambitions are more modest– we're going to classify Iris flowers based solely on the length and width of their sepals and petals.

The Iris genus entails about 300 species, but our program will classify only the following three:

- Iris setosa
- Iris virginica
- Iris versicolor

2 Dataset

Fortunately, this being our starting problem for machine learning we don't need a very huge amount of data. We require training datasets and test datasets. There is a set of 120 entries for this problem which we will use. Refer Dataset.

There are five columns in this dataset. The first four are feature and and the last column is the label. Each label is naturally a string (for example, "setosa"), but machine learning typically relies on numeric values. Therefore, someone mapped each string to a number. Here's the representation scheme:

- 0 for Setosa
- 1 for Versicolor
- 2 for Virginica

3 Models and Training

The Iris classification problem is an example of supervised machine learning in which a model is trained from examples that contain labels.

A **model** is the relationship between features and the label. For the Iris problem, the model defines the relationship between the sepal and petal measurements and the Iris species.