What happens when there are too many features in a dataset?

If we want to develop some kind of prediction or classification model (such as a neural network), it can take a lot of time for the model to run. If there are 60 or 70 variables in our dataset, that can be very time-consuming and also sometimes reduce the accuracy.

In this situation, we use Boruta for feature selection.

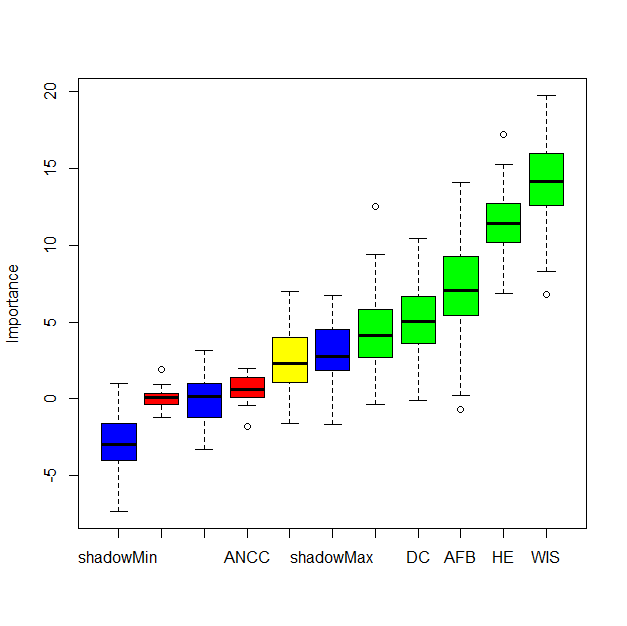
**Boruta**

* *God of Forest* (Mythological god)
* Based on Random Forest

**How it Works:**

Suppose you have 60 attributes in your data. For each attribute, it creates a **shadow attribute**. The shadow attribute is all the variables shuffled across so that there is randomness. Then we create a classification model with the **shadow and original attributes**. And assess the importance of all attributes.

**The idea is**:  
If the variable is not doing better than the shadow attribute in terms of importance, obviously we should **not have** that kind of attribute in our model.



The Boruta plot shows feature importance using Random Forest. Green boxes (WIS, HE, AFB, DC) are confirmed important features. Red (ANCC) is rejected and should be excluded. Blue boxes are shadow (random) features used for comparison. You should keep the green features and drop the red ones for modeling.