Seeing the Wood for the Trees: Interrogating the Structure of Random Forests

Chris Harbron

Discovery Statistics

AstraZeneca

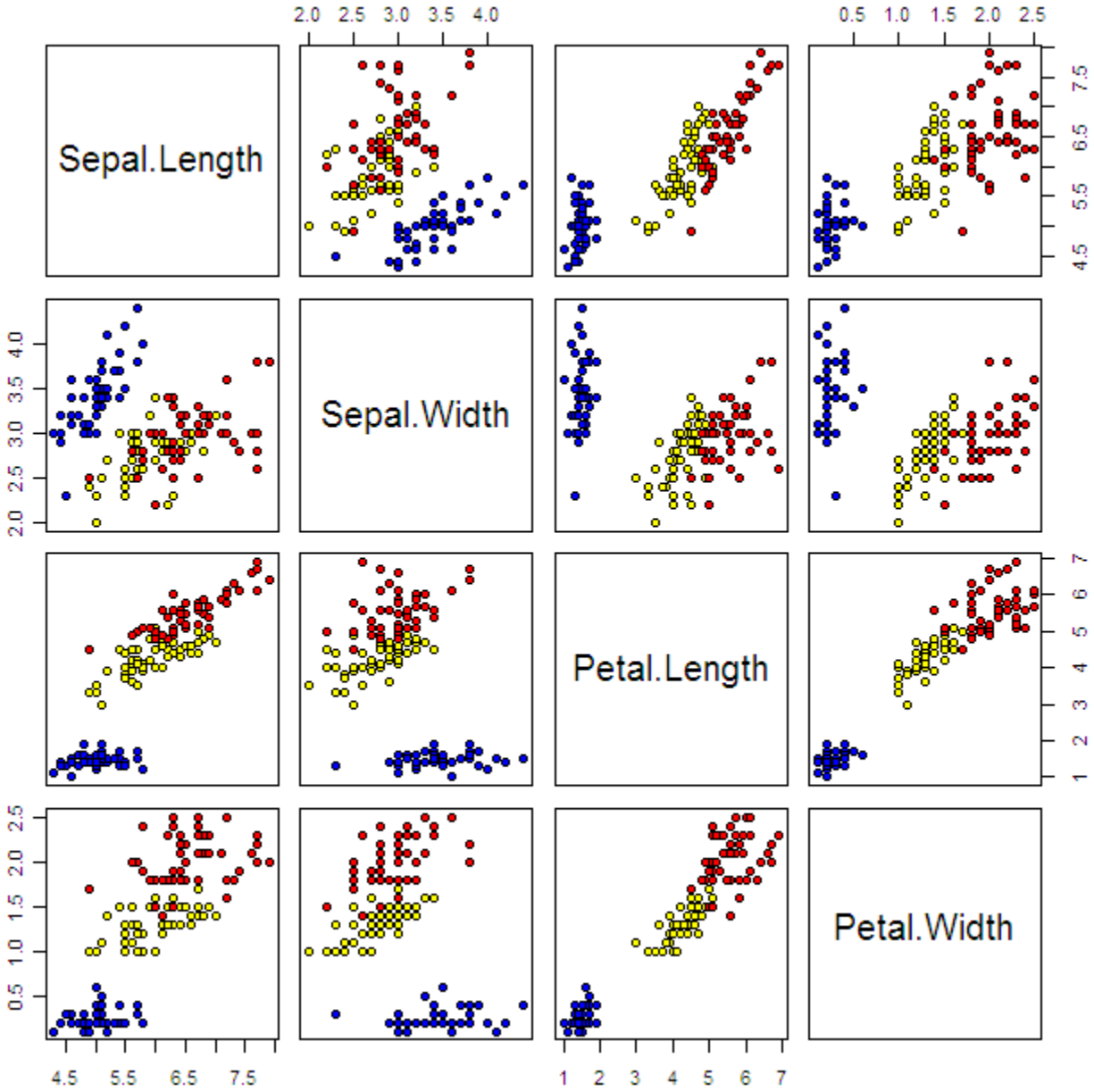
AstraZeneca
life inspiring ideas

Random Forests

- First published by Breimann and Cutler in 2001
- Popular modelling tool for classification, regression and survival analysis with (highly) multivariate data
- Powerful, robust, easy to apply with many attractive properties
- Breimann was passionate that modelling was about more than just predicting:
  - “With scientific data sets more is required than an accurate prediction”
  - “Looking inside the black box is necessary”
  - Variable Importance Measures, Proximities
- SURF : An extension of Variable Importance Measures

Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Fisher's Iris Data



Setosa

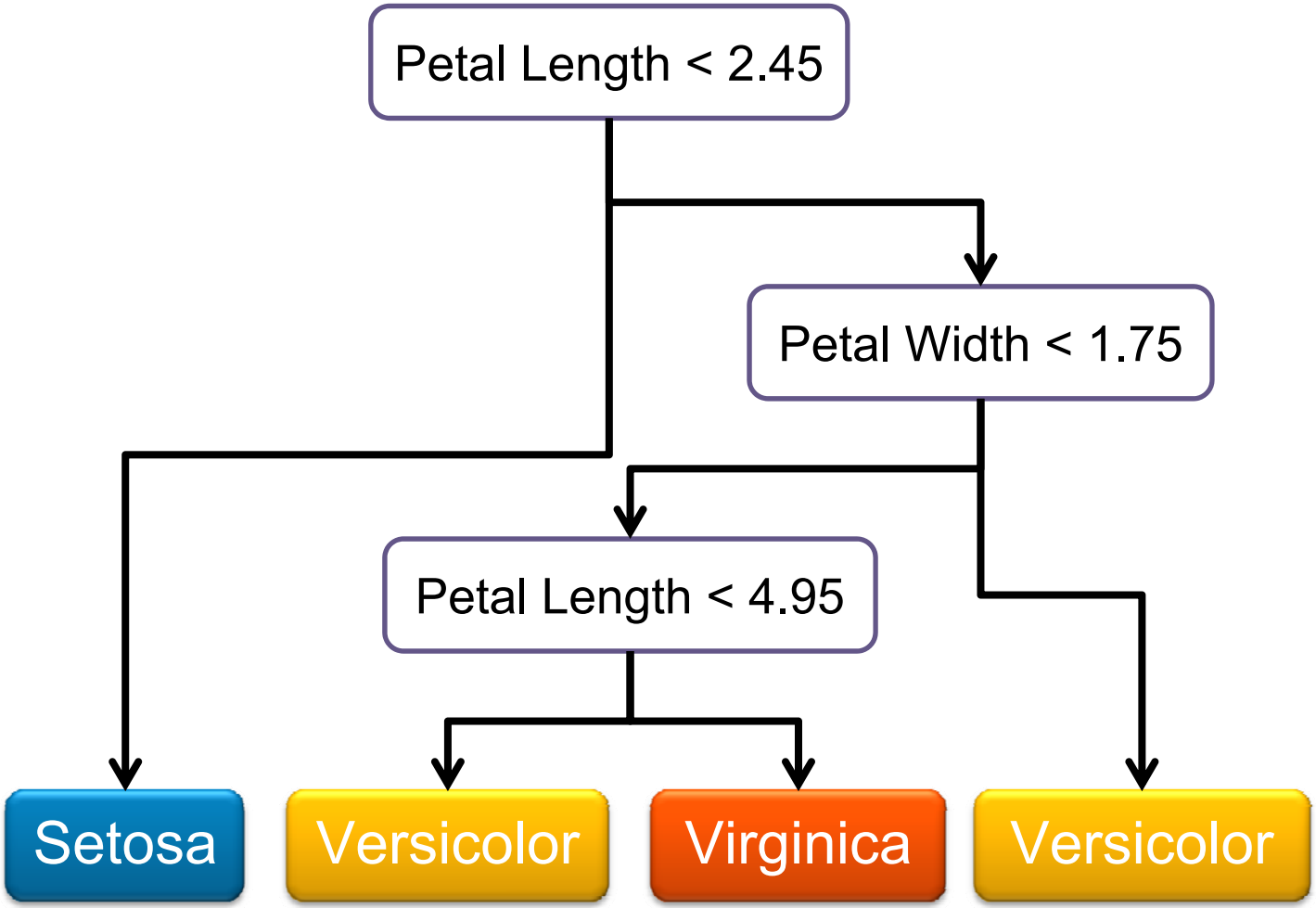
Versicolor

Virginica

AstraZeneca Discovery Statistics

Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

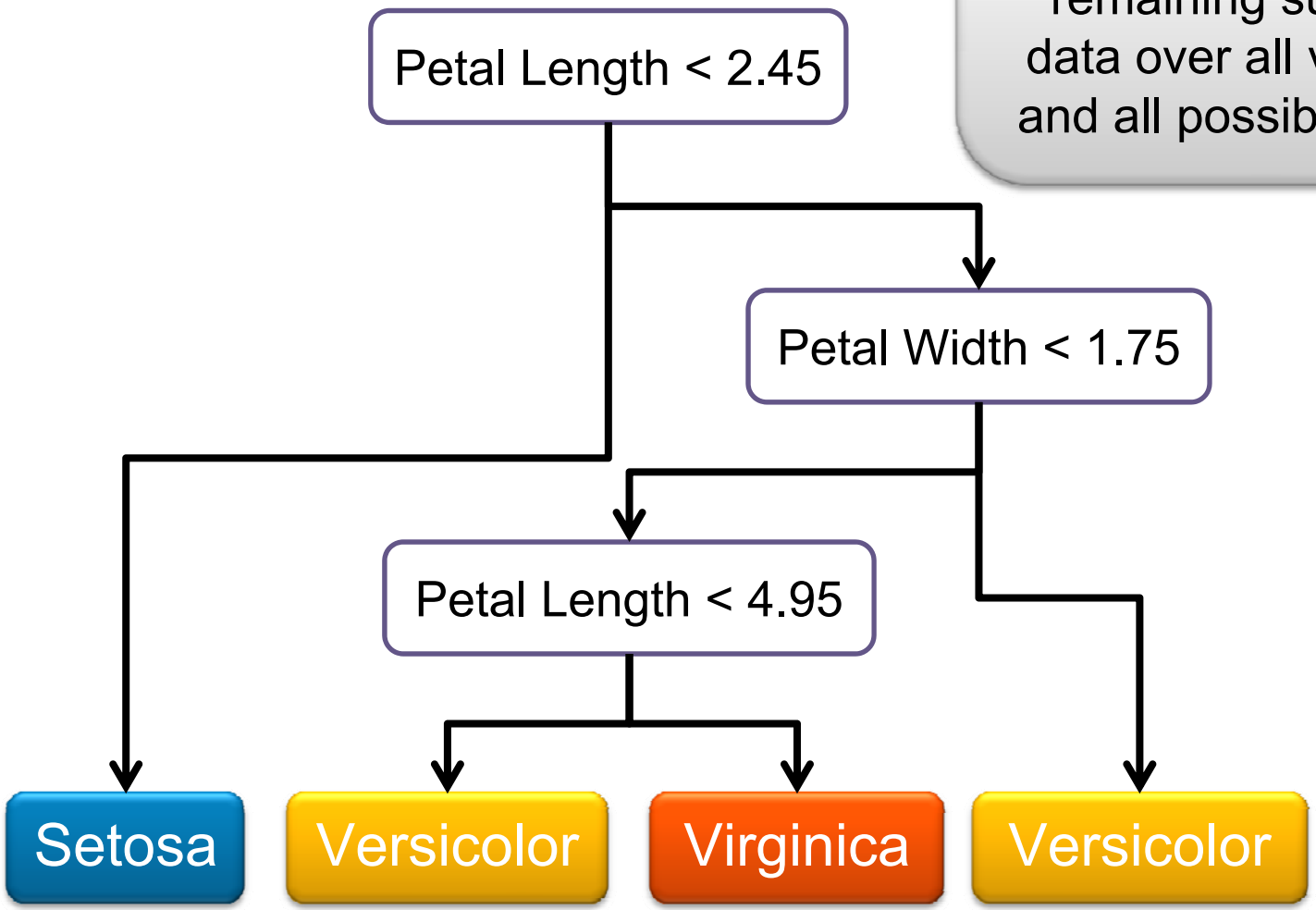
Decision Trees



Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Decision Trees

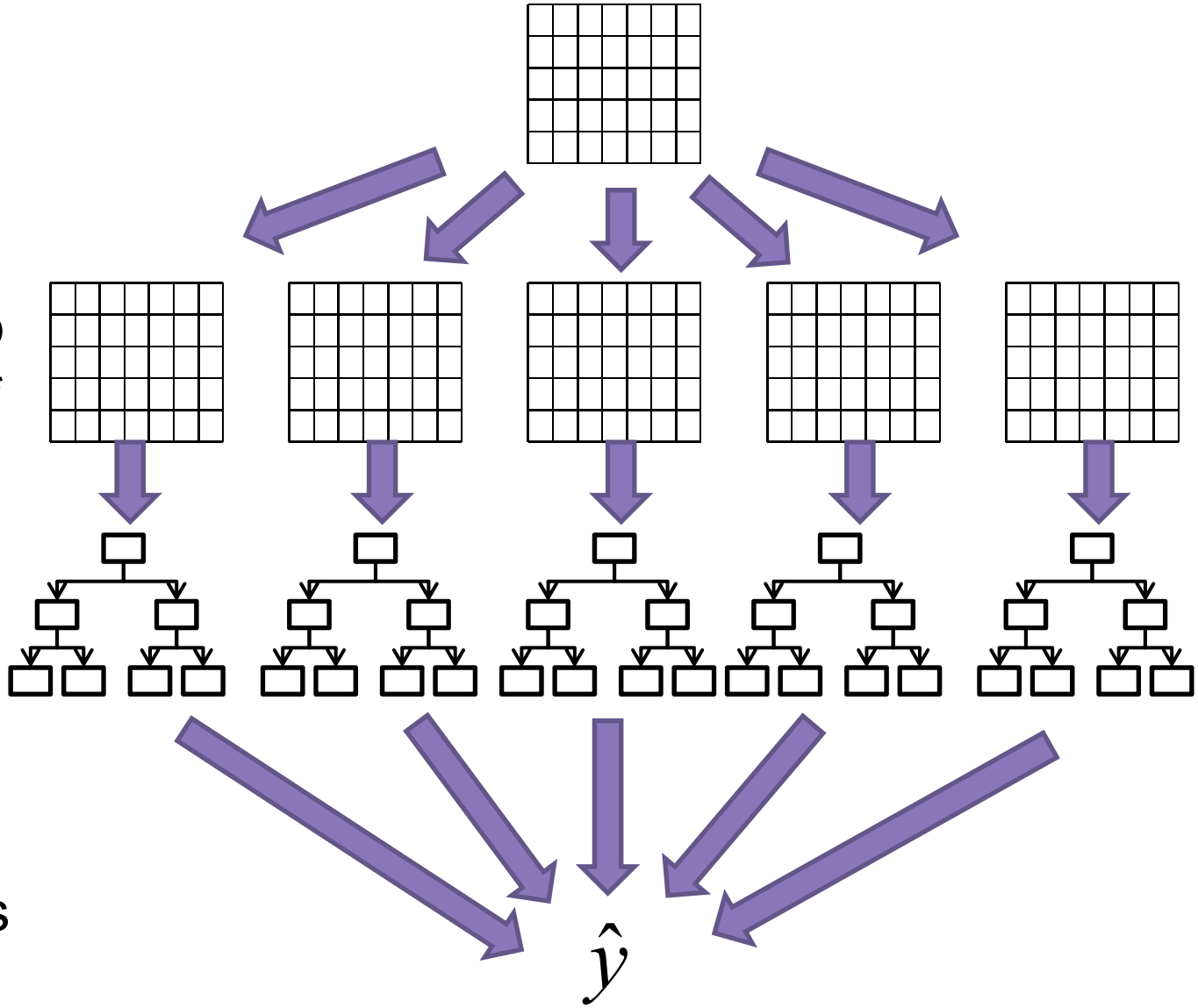
At each split, maximise the improvement in prediction with the remaining subset of data over all variables and all possible values



Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Random Forests

Lots of trees with 2 perturbations* of the data



Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

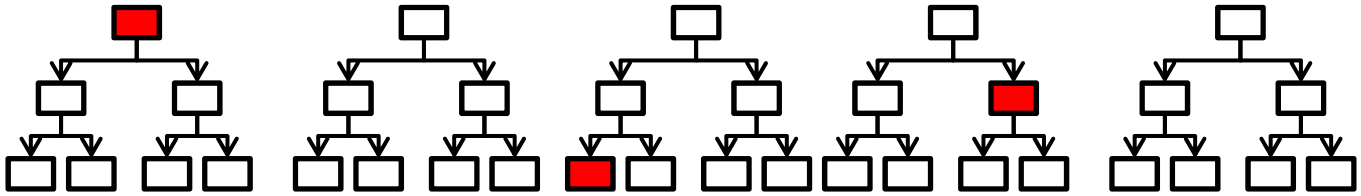
Variable Importance Measures
2 Standard Measures

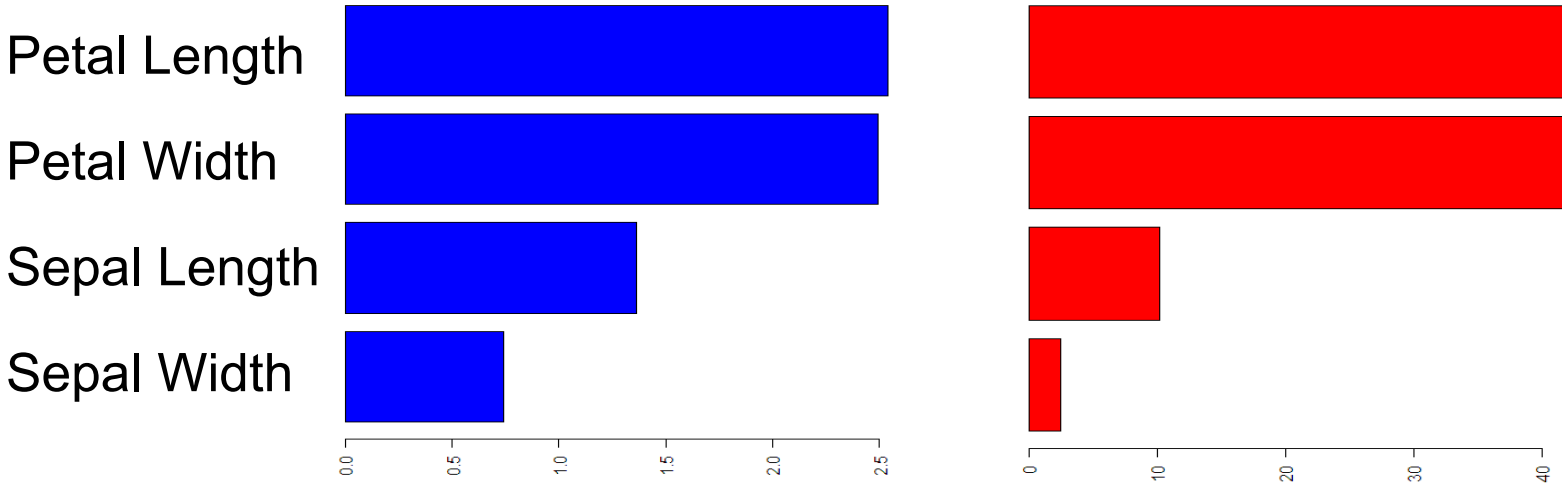
Permutation

- Change in accuracy of predictions when permuting each variable in turn

Reduction in Impurities

- Sum of the reduction in Gini index / MSE over all splits by that variable in the forest





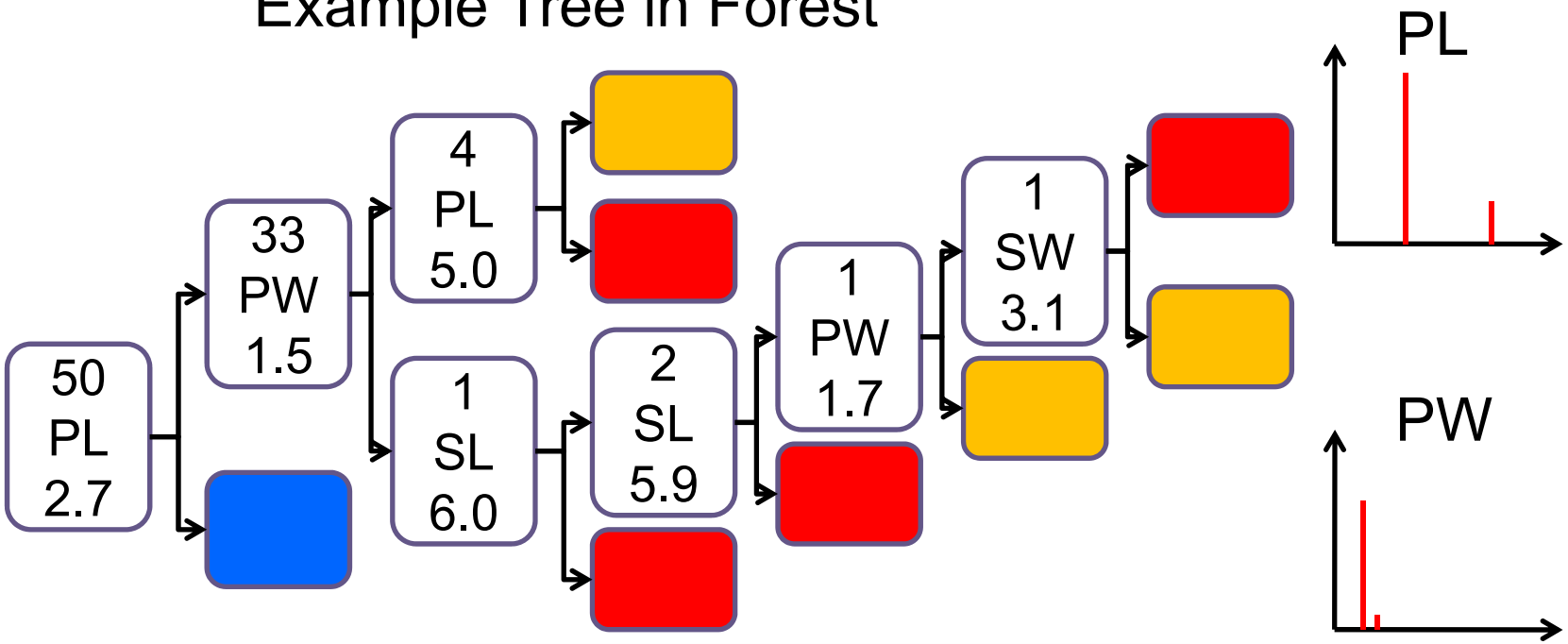
Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

SURF :
Understanding Splits In Random Forests

Split Total Gini Score By Predictor Variables

Split Total Gini Score By Predictor Variables and their Split Values

Example Tree in Forest

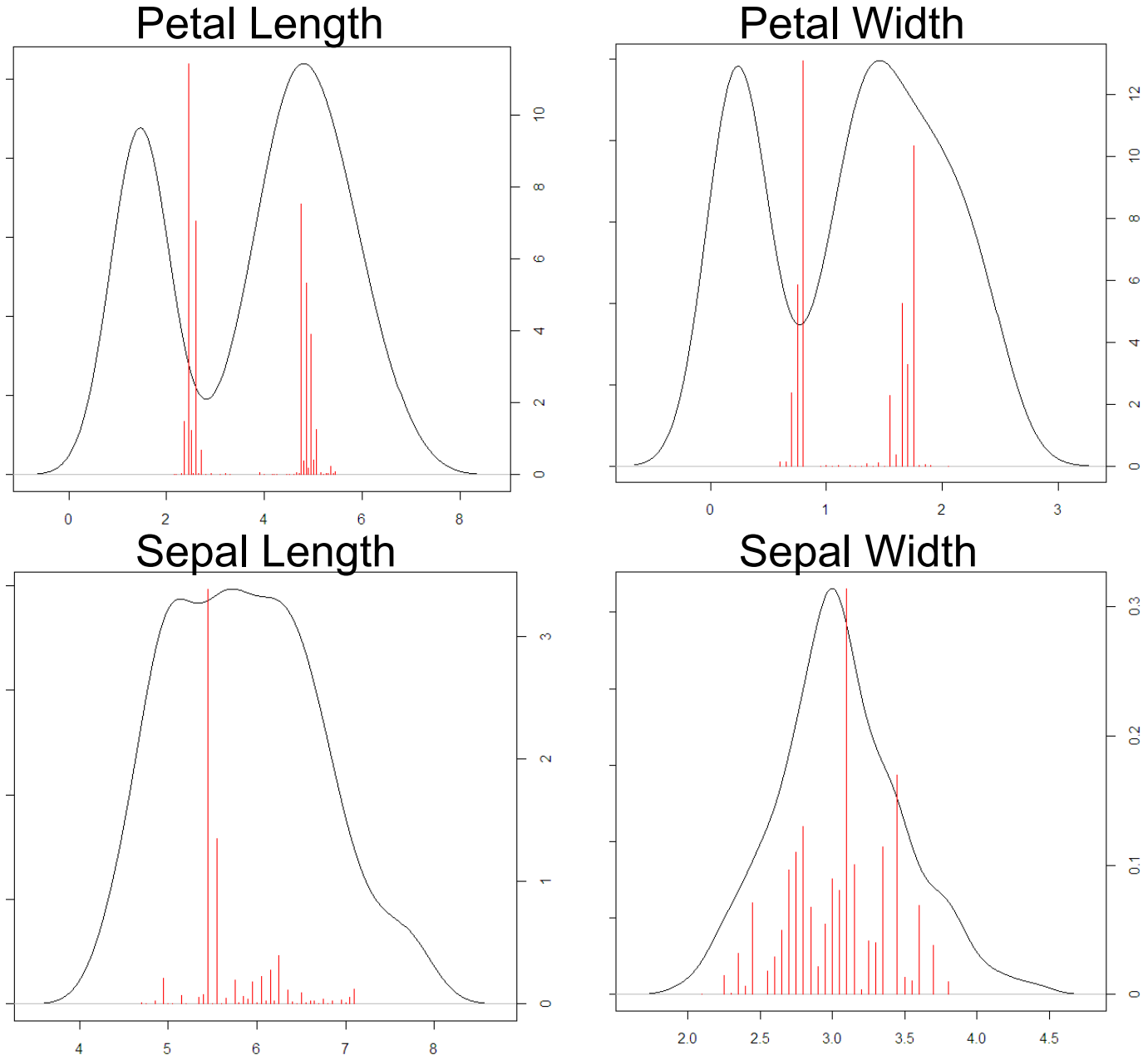


∑ Sum Over All Trees

Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

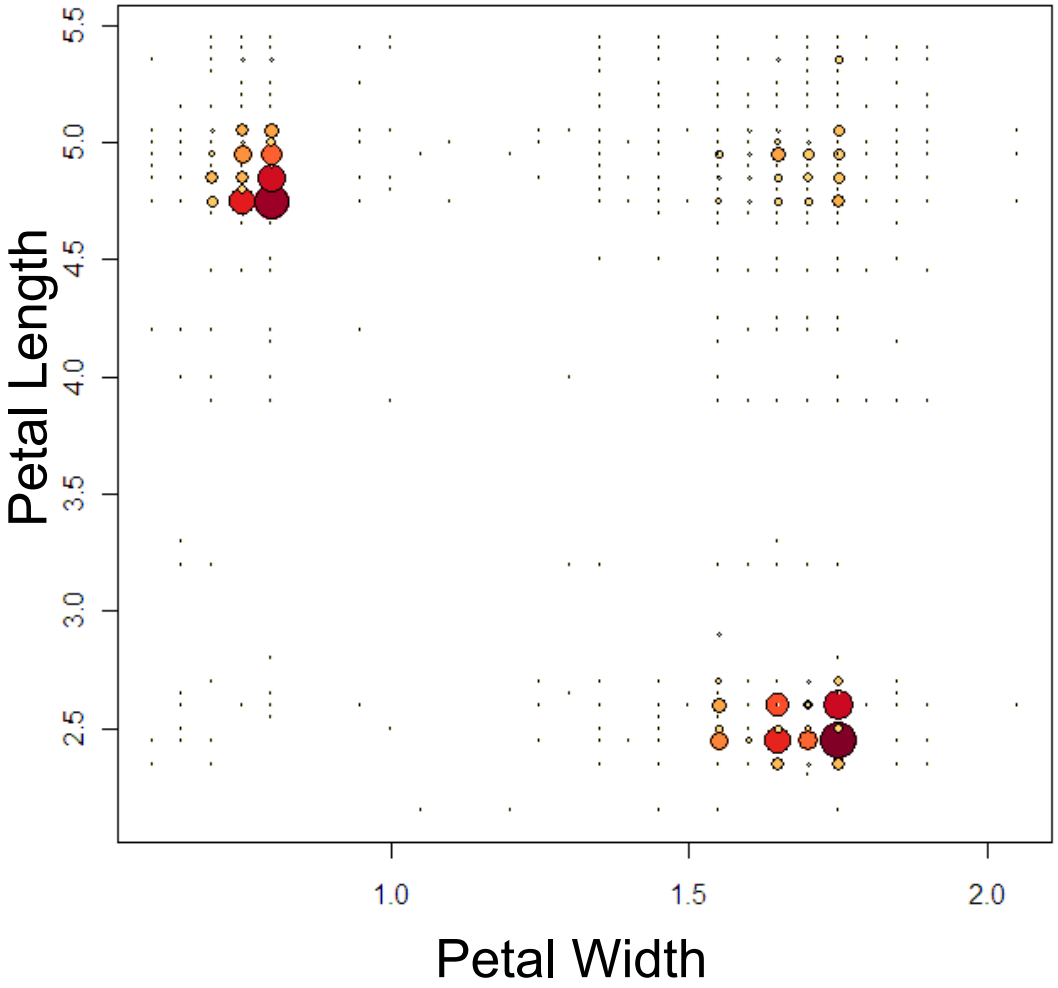
AstraZeneca Discovery Statistics

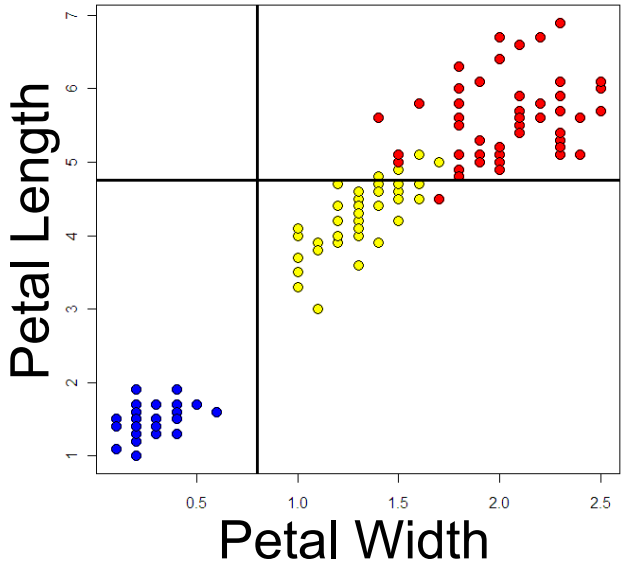
Univariate Results

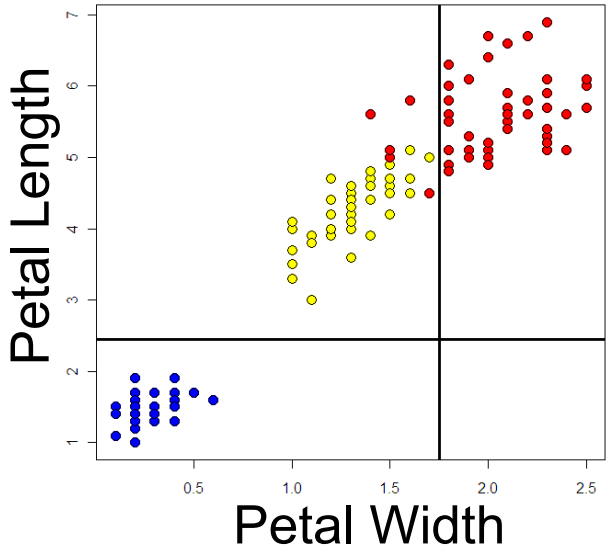


Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Pairwise Results



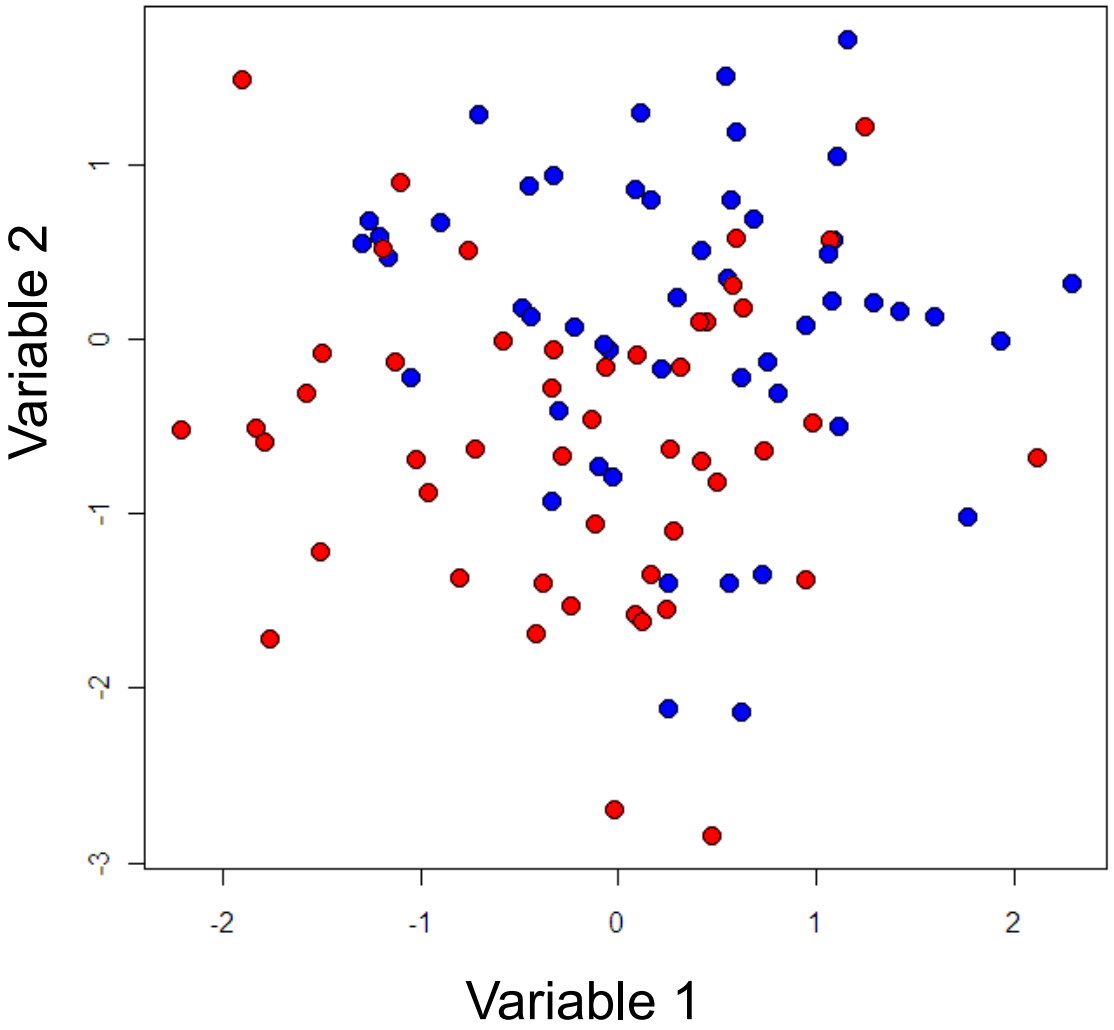




AstraZeneca Discovery Statistics

 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Simulated Data
Continuous Response



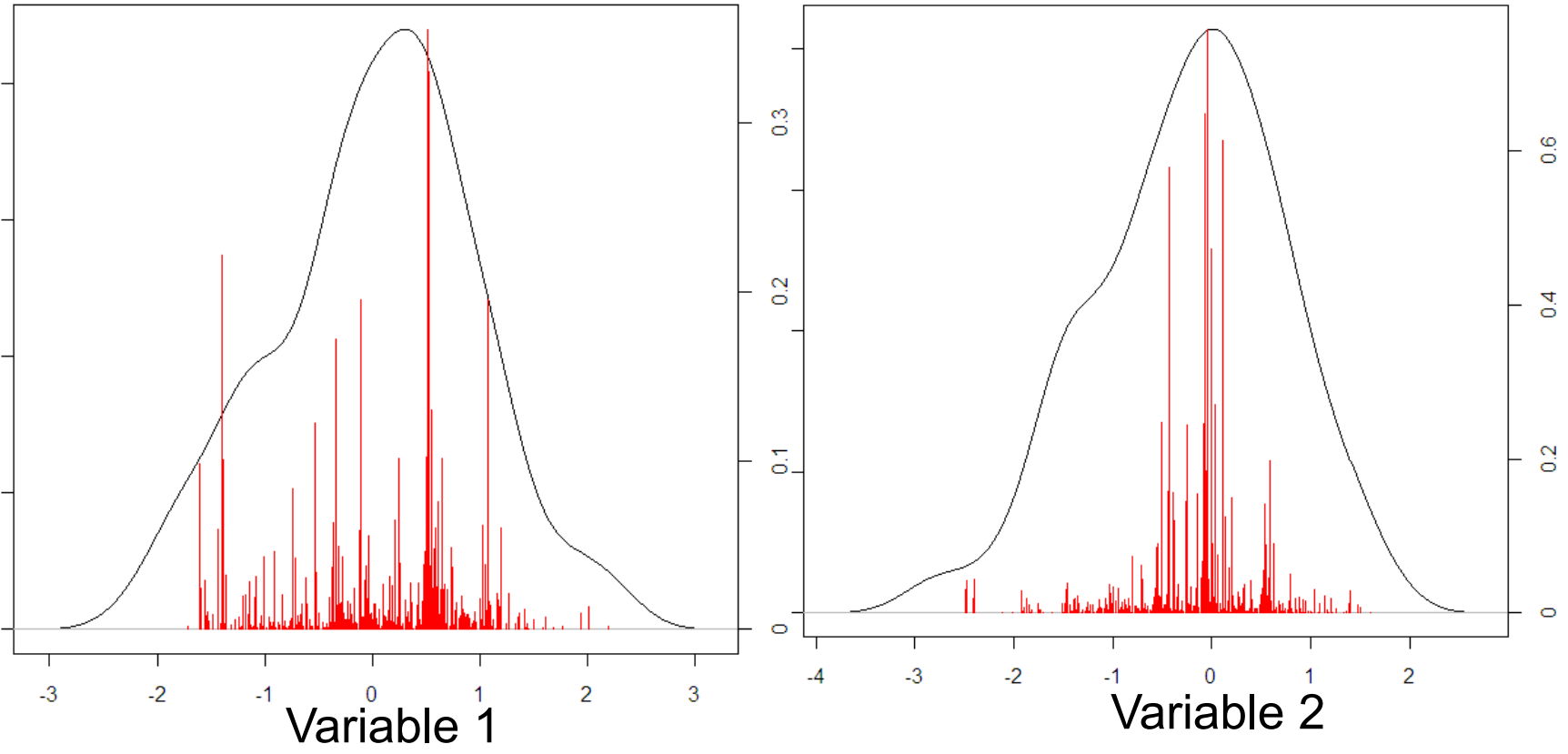
$$y \sim Bin(1, p)$$
$$Logit(p) = Var1 + Var2$$

Four additional non-correlated noise variables included in model

AstraZeneca Discovery Statistics

 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

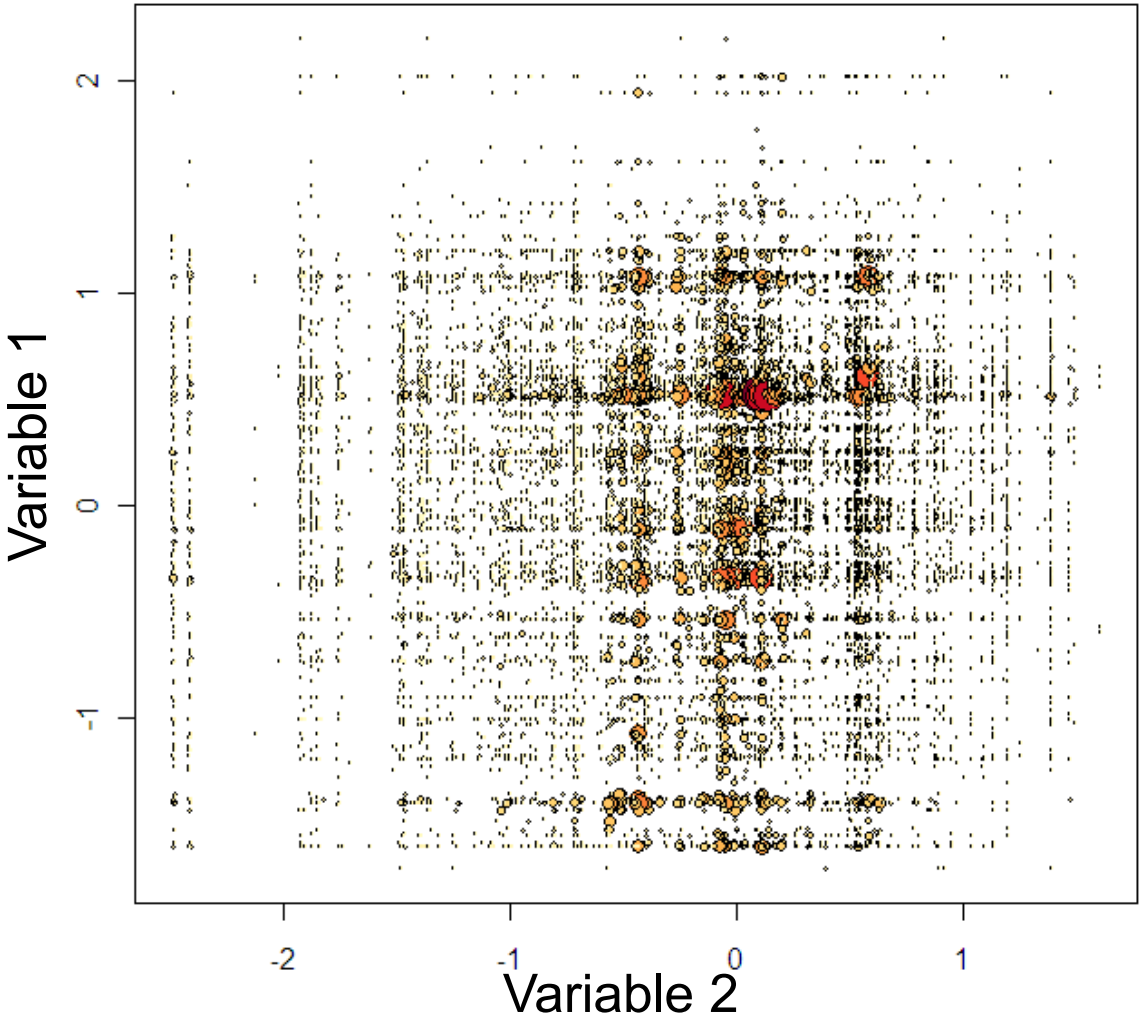
Simulated Data



Splits spread over a much more diffuse set of values

 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Simulated Data

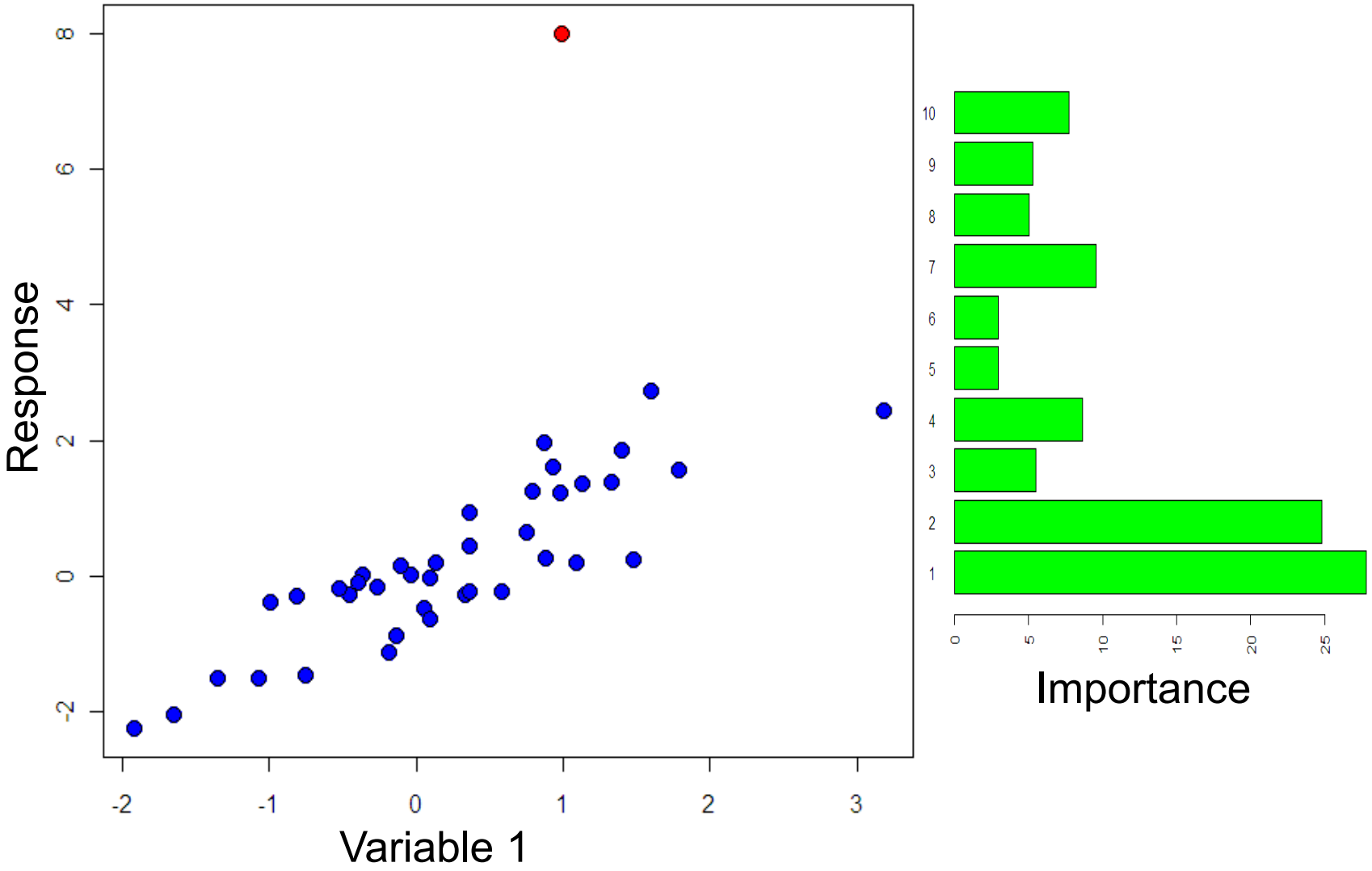


Splits spread over a much more diffuse set of values

Splits in the 2 variables acting independently from each other

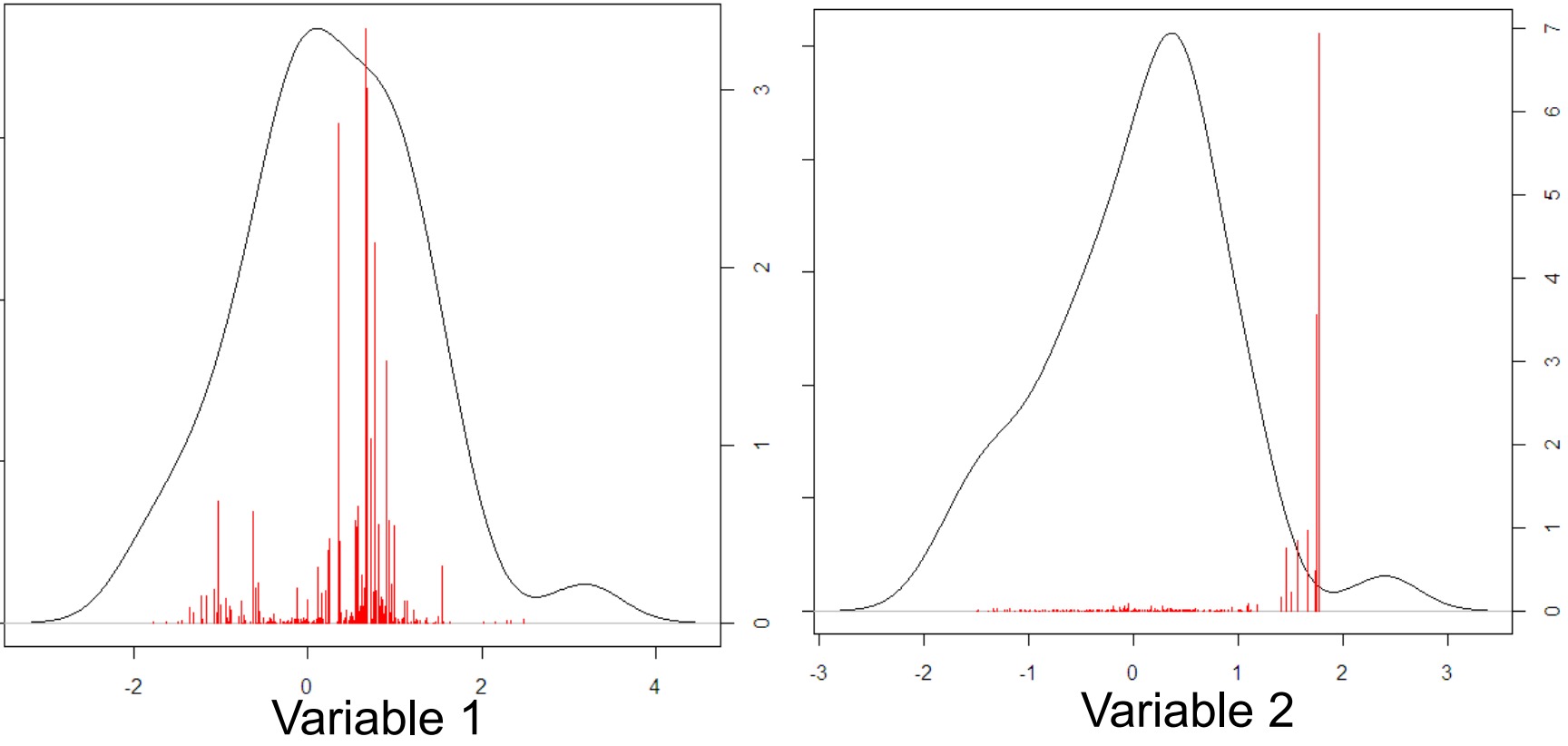
 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Simulated Data - Outliers



 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Simulated Data - Outliers



Diffuse set of splits centered around median

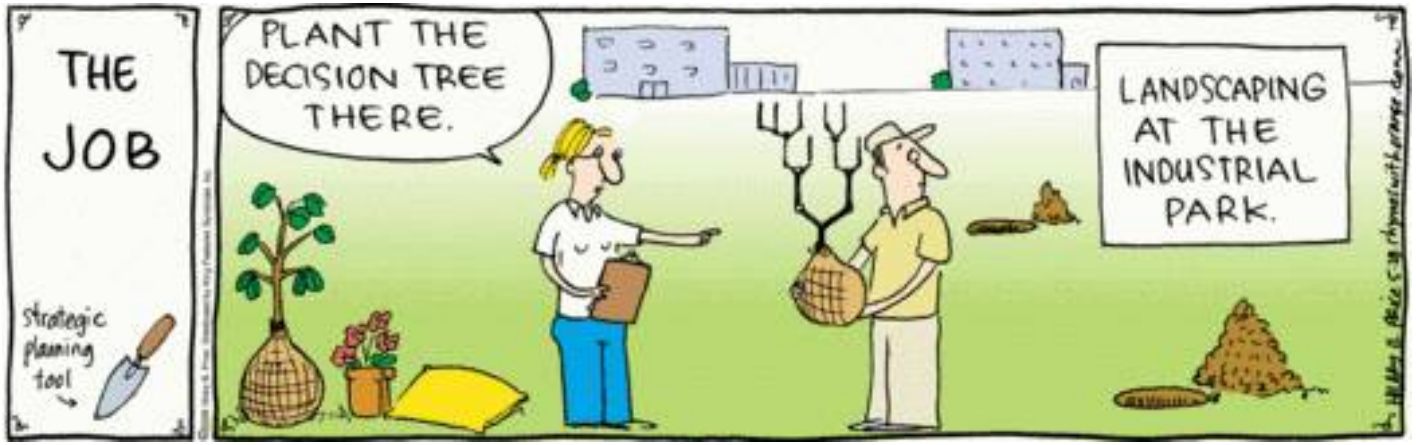
All splits separating single outlying value

AstraZeneca Discovery Statistics

 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010

Summary

- Random Forests are a powerful and robust multivariate modelling technique
- Beyond acting as a black-box predictor, random forests give insights into the underlying structure of the data through variable importance scores
- This can be taken a step further by considering the locations of the splits
- Also gives a valuable insight into the quality of the model



 Chris Harbron, Seeing the Wood for the Trees: Interrogating the Structure of Random Forests, NCSC 2010