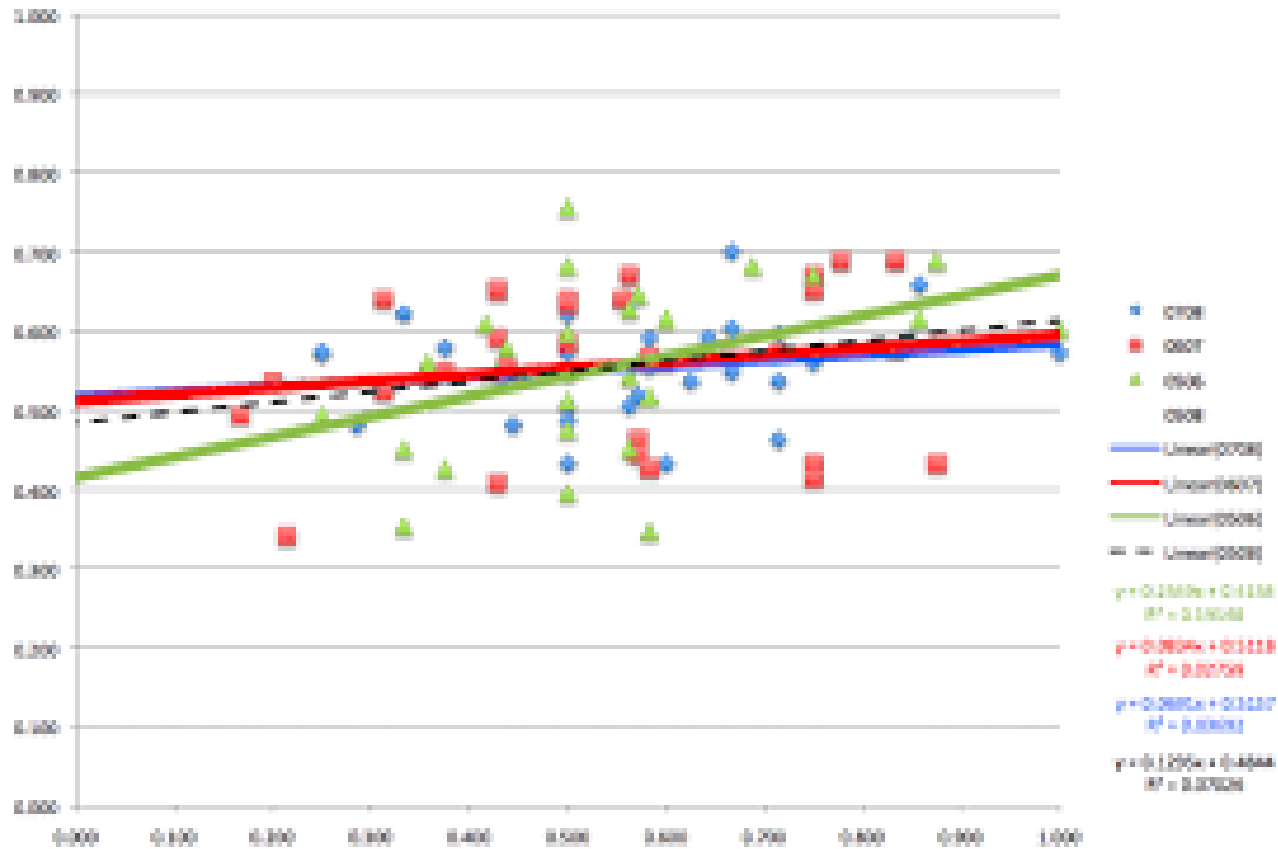


Can't see the trees for the forest

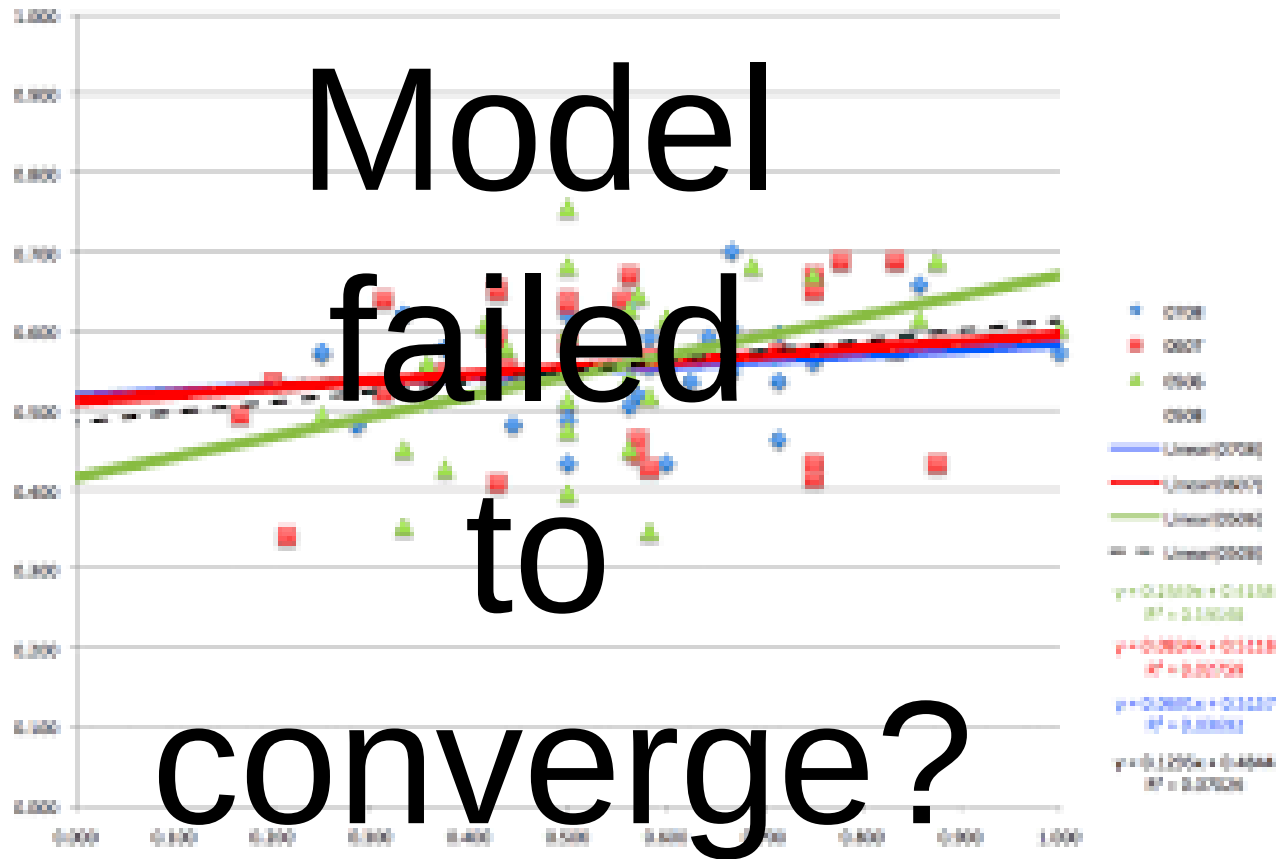
28th June 2019
Emma Mills



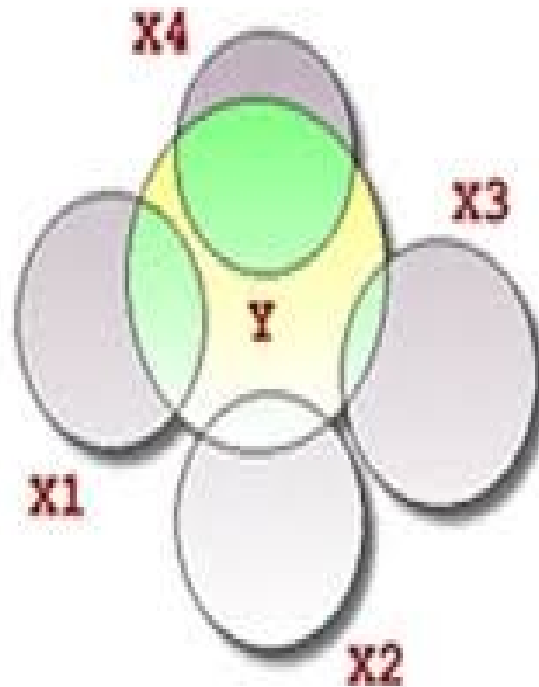
Up - Sell



Up - Sell



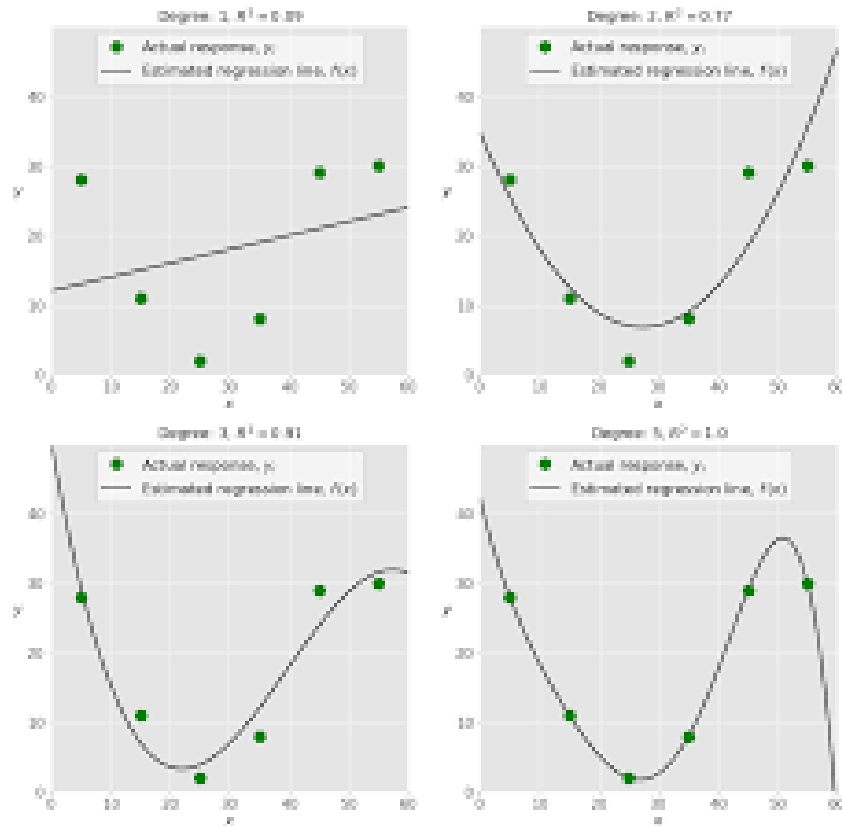
Up - Sell



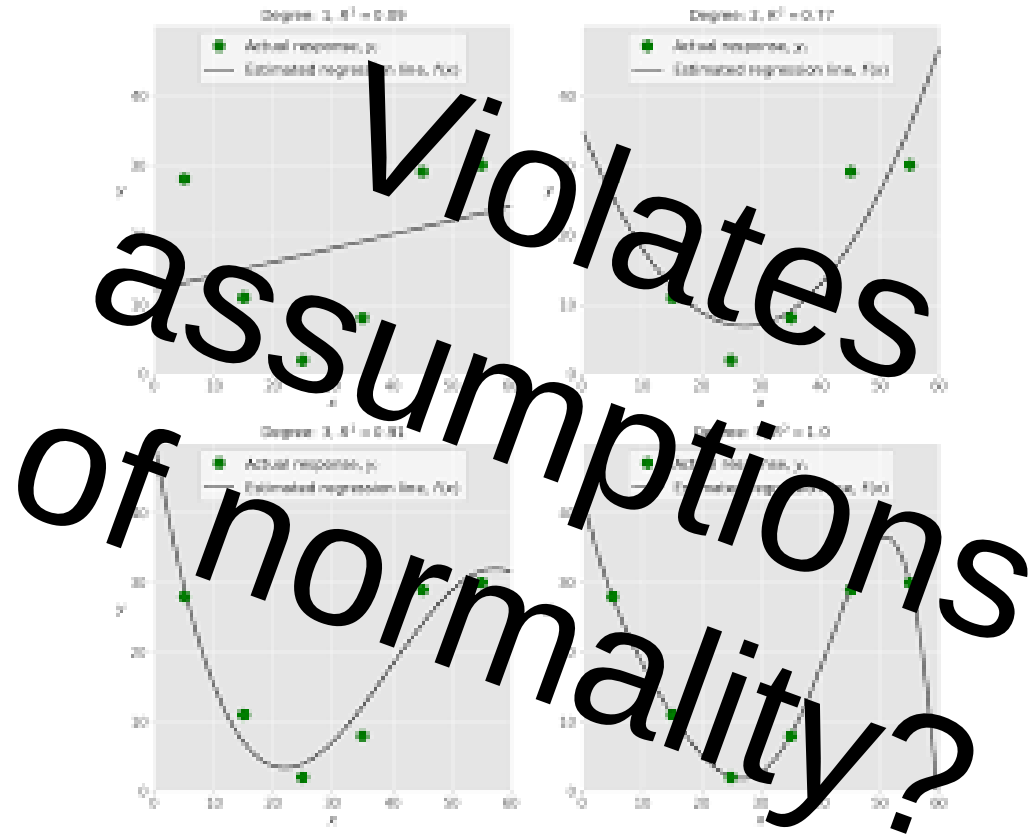
Up - Sell



Up-sell



Up-sell



Consider...

Random Forests

- powerful
- allows multicollinearity
- linear and non-linear effects within the same analysis¹



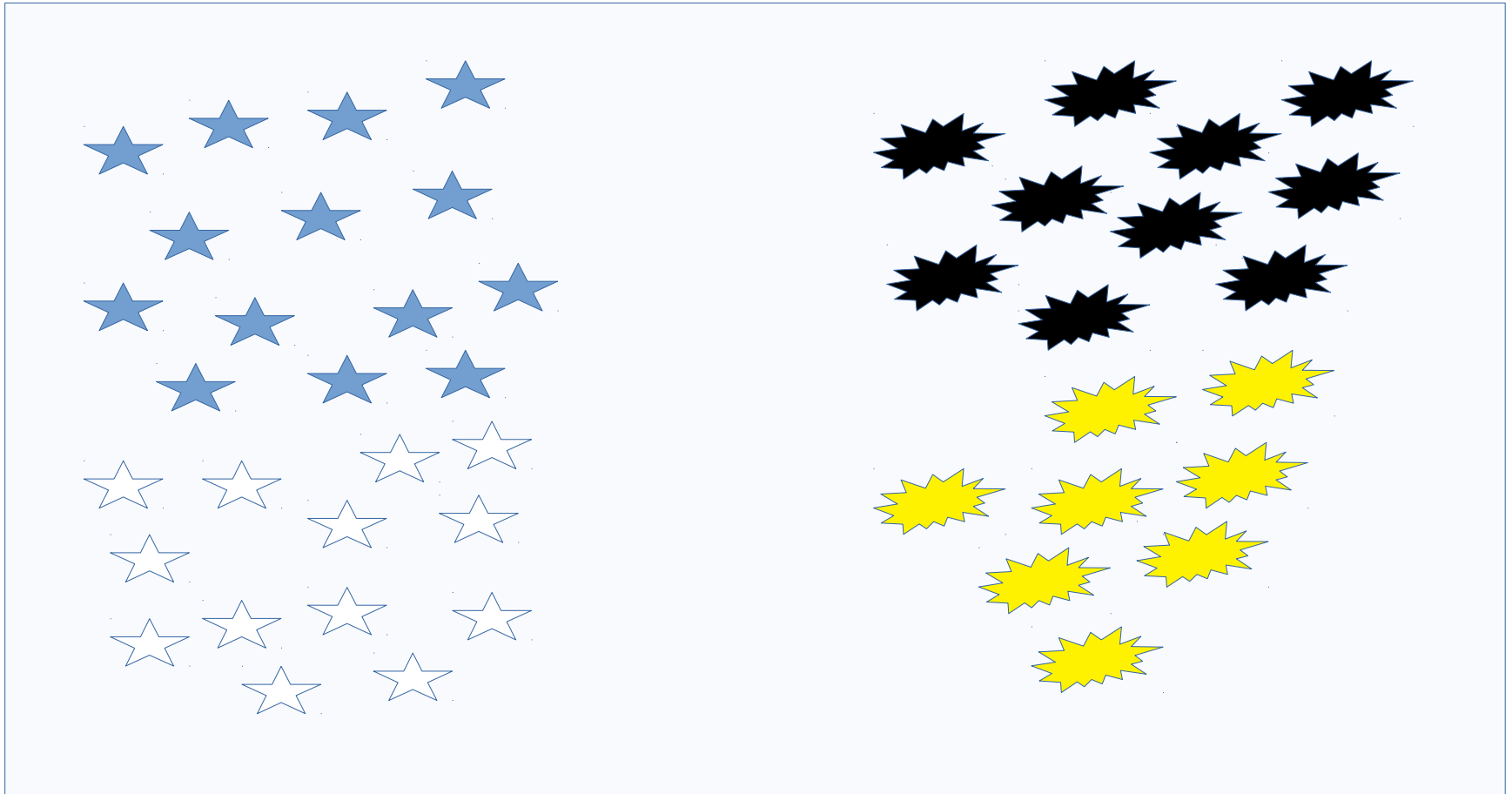
Additional benefits:

- Works with a small sample size
- Insensitive to order effects
 - Sampling process mitigates against experimenter bias
- Can assist with variable reduction – if required²
 - More stable than stepwise regression³
 - Works with observed variables rather than latent variables⁴

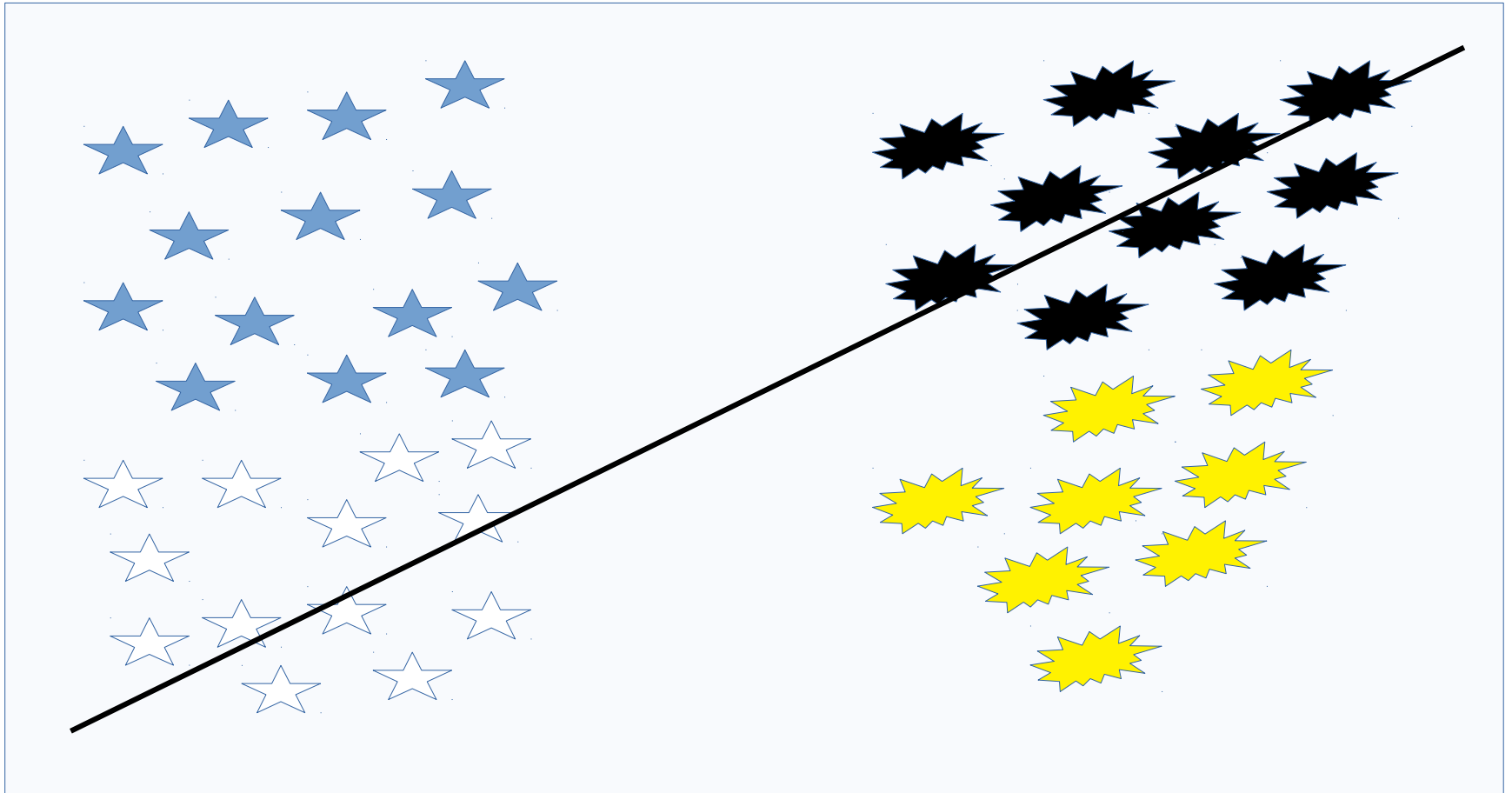
Health warnings:

- Black Box – a lot of stuff under the hood that you don't mess with unless you *really* know what you're doing!
- Not a replacement – more of a complimentary tool
- Data driven method – your data sample may not align with your theoretical expectations
- Variable selection takes a lot of time

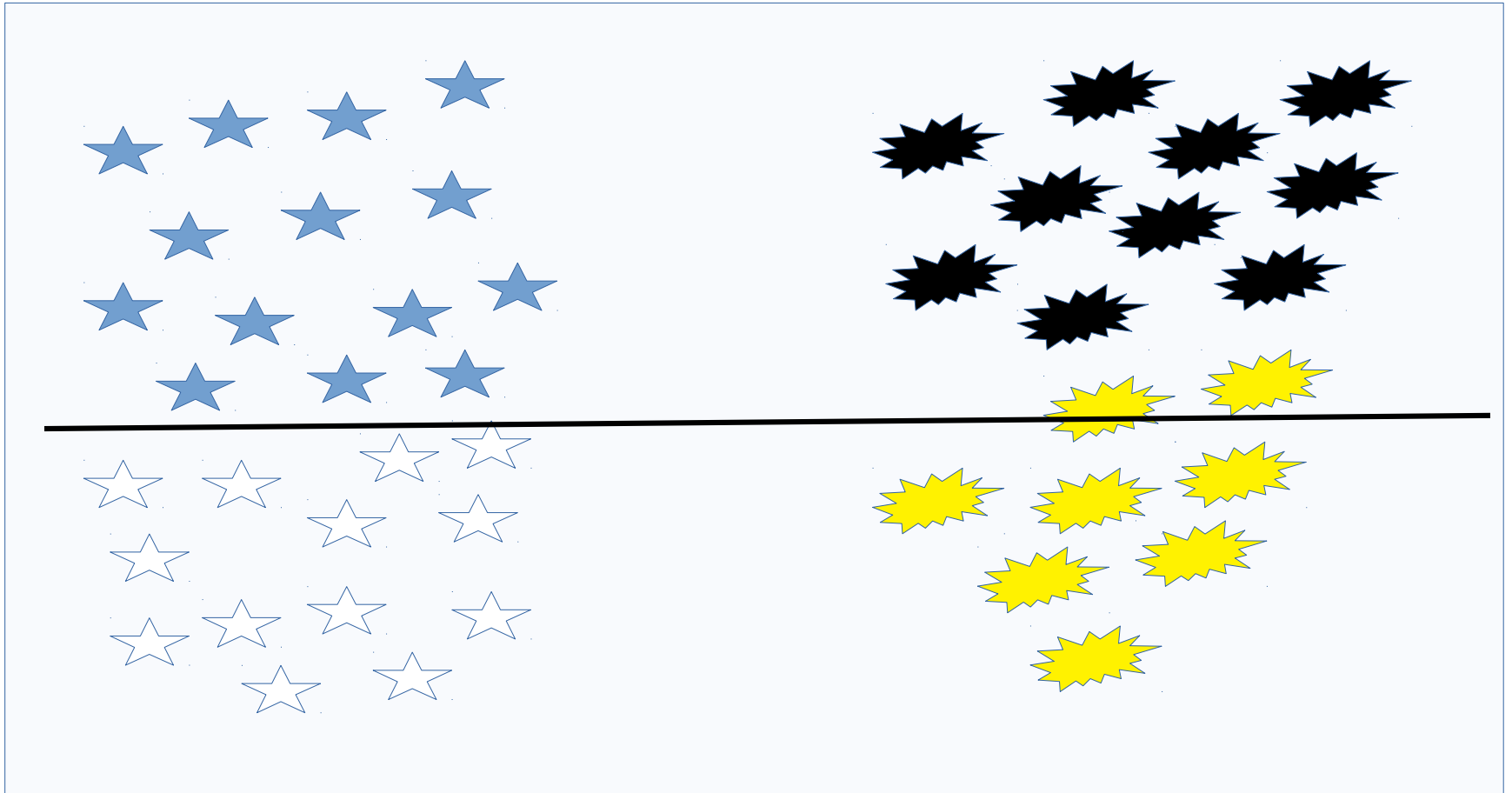
How does it work? A simple example



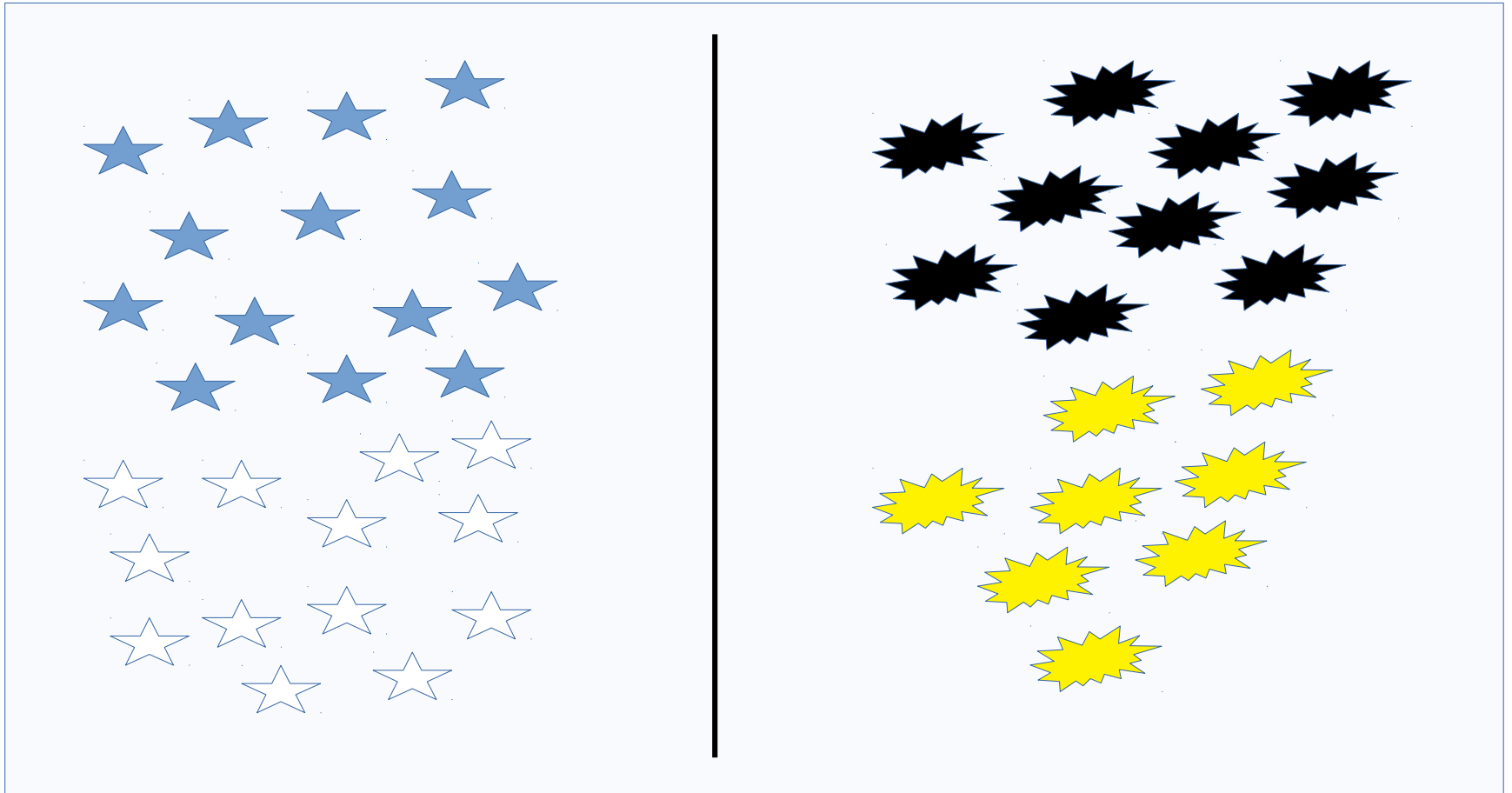
Random splits across the data set by one variable



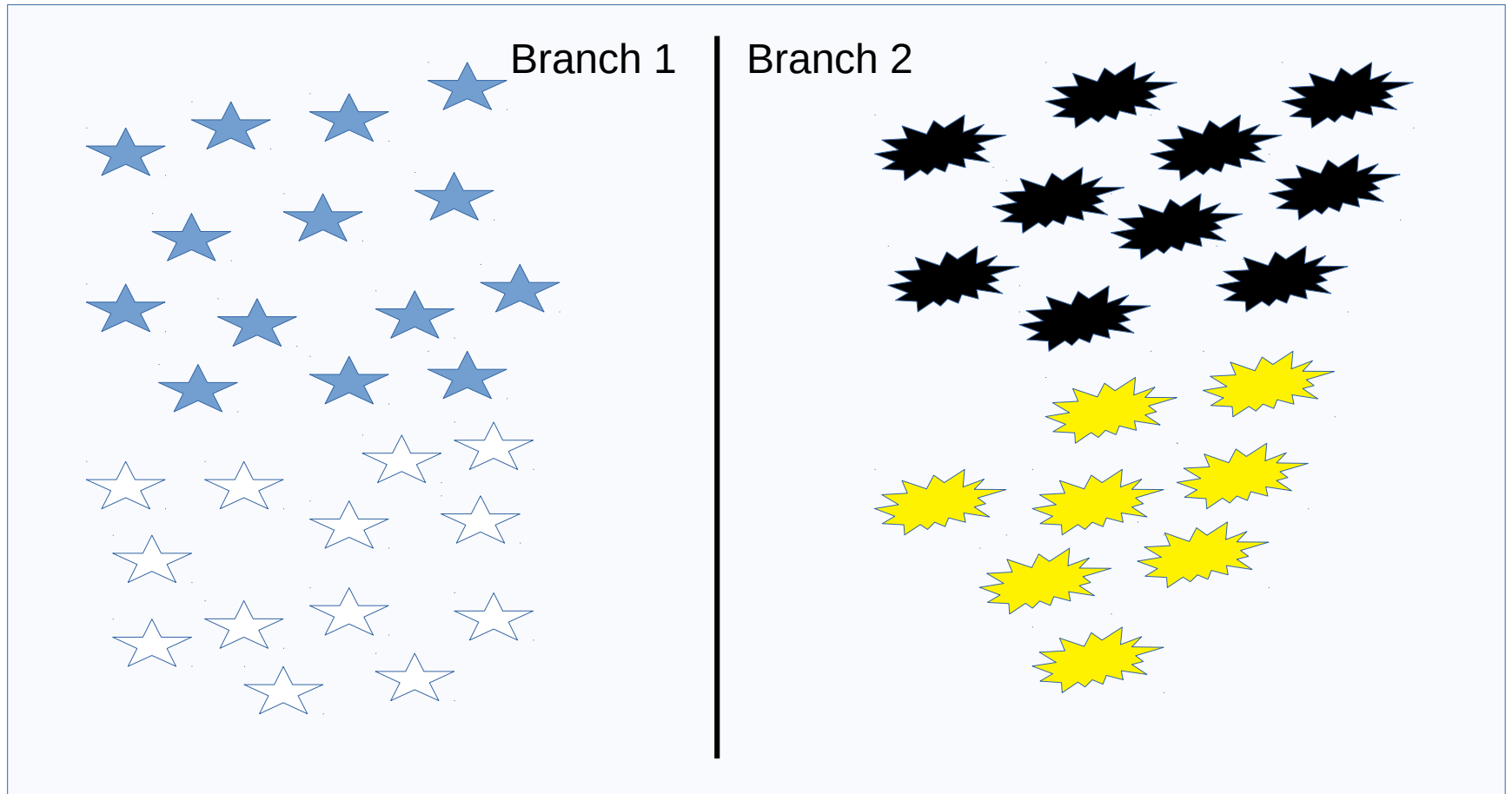
Recursive partitioning occurs until...



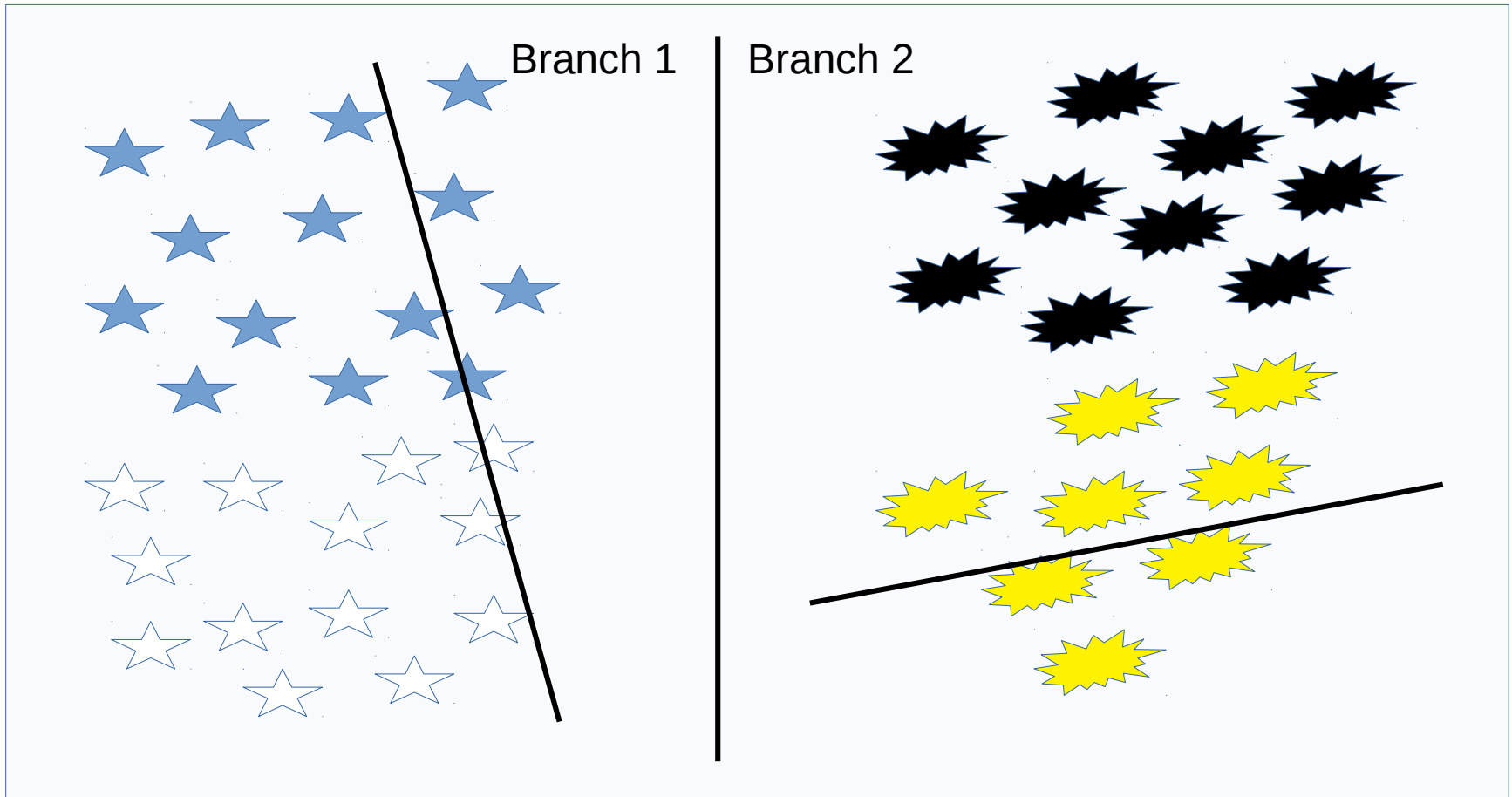
Impurity reduction



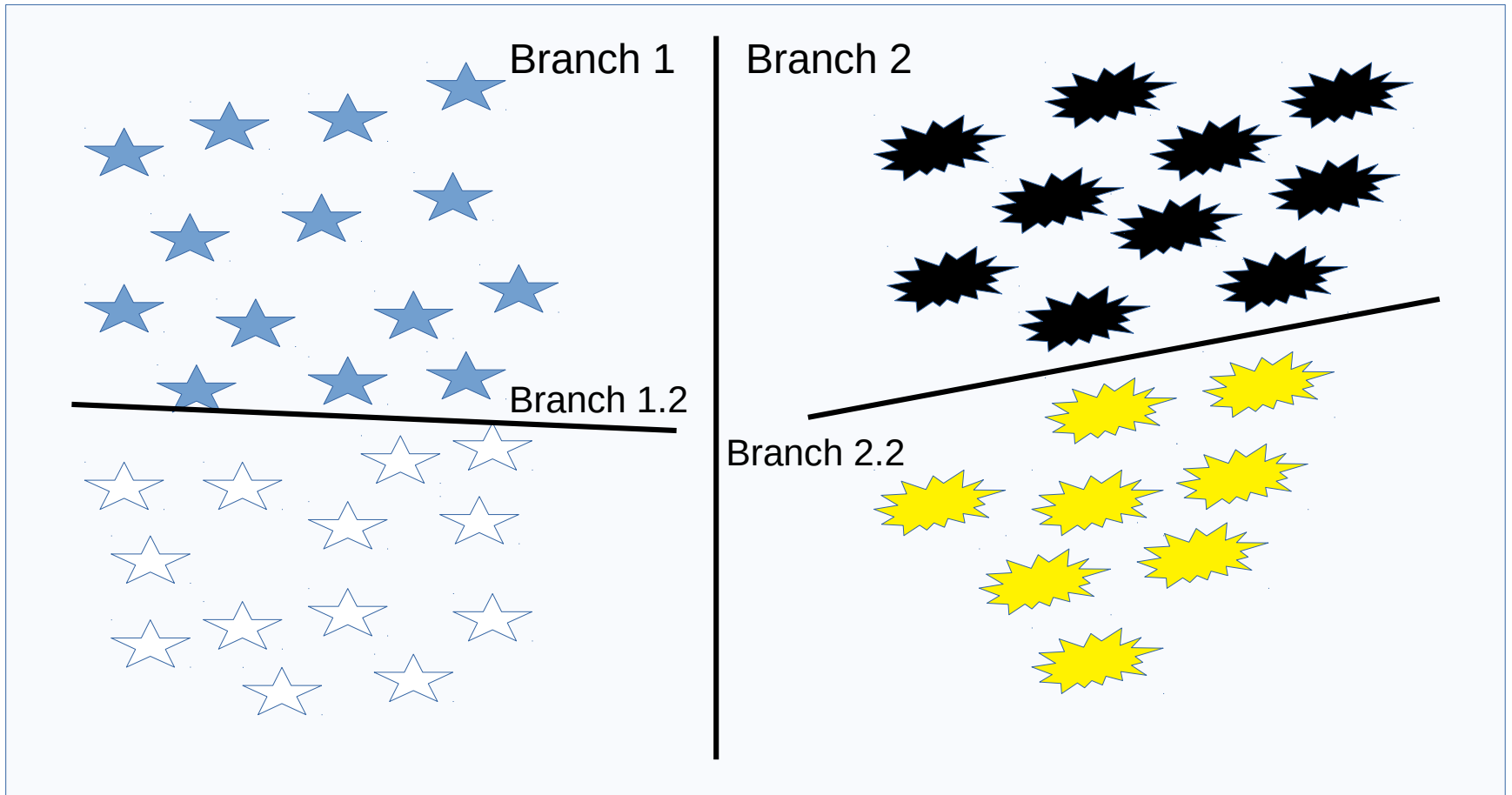
Binary split...



And repeat...by another variable



And repeat...



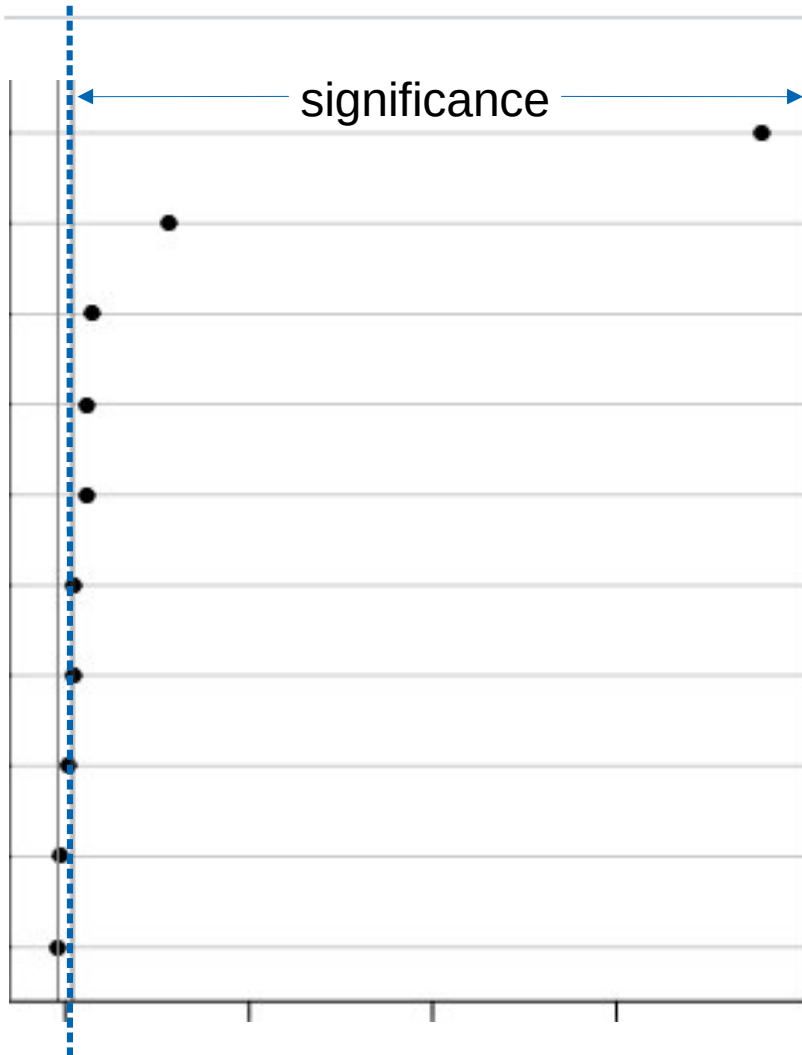
Random Forest

- One such tree is constructed many times with random sampling
- Random sampling without replacement
 - Across predictors
 - Different predictors as initial split
 - Across observations

Results

- Aggregation over all the trees: response variable with the most vote wins
- Variable importance needed though
 - Sampling means not all predictors are considered across all trees
 - Assessed by assigning new levels to variables and testing for an effect - a kind of sensitivity analysis but within and across variables
 - By product = cross validation: training and testing across full sample each time

Variable Importance Graphic



- Points to the right = significant variables
- Advice for variable selection is to *exclude* variables that are within the same range as *negative* variables
- Re-write formula with **identified variables**

Figure taken from Tagliamonte & Baayen 2012

Example plot

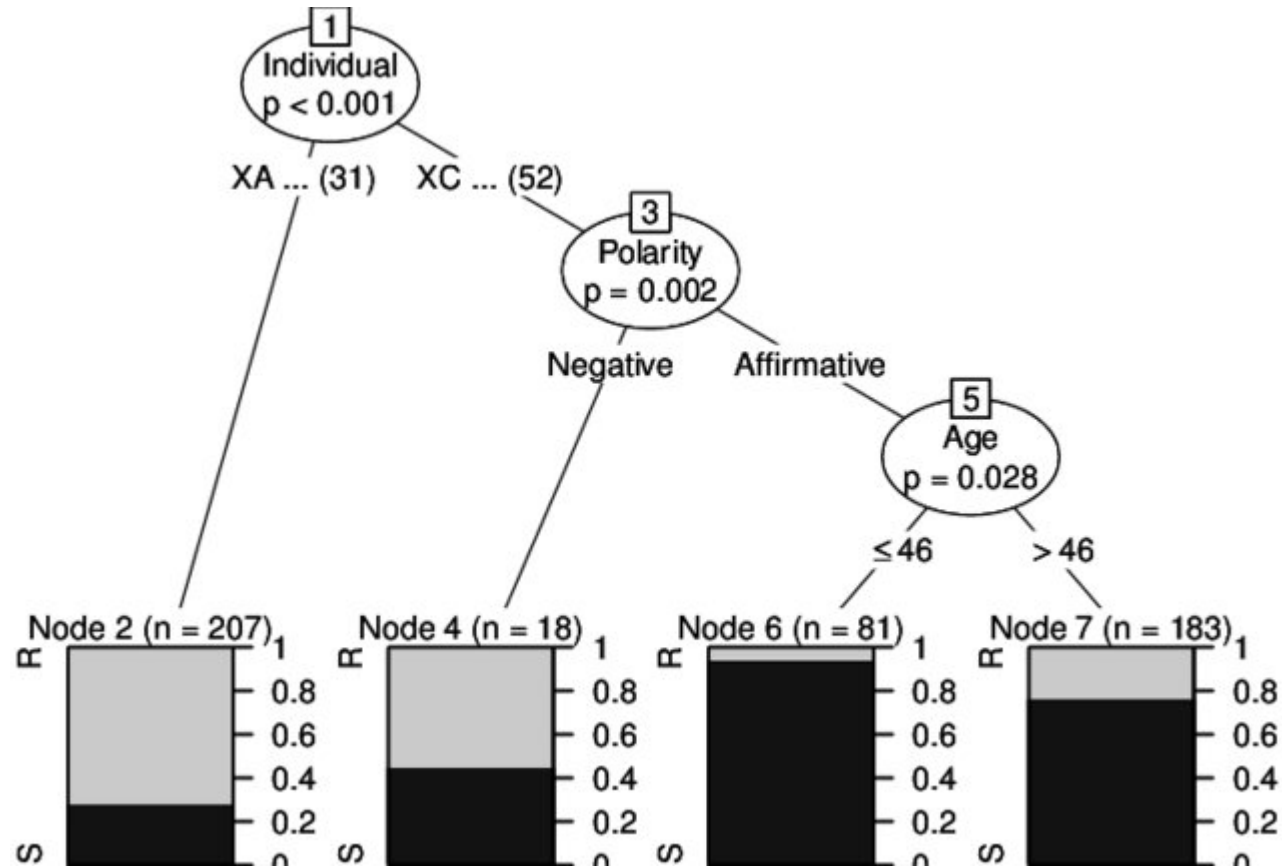


Figure taken from Tagliamonte & Baayen 2012

Conclusions

Random Forest offers

- Easy to implement method – formula just like regression
- Very flexible – small, non-linear sample – no problem
- Sampling method that validates itself in the process
- Variable selection routines if required
- Plots are quite intuitive to interpret