# 1 Pruning the tree

Pruning is the process of removing leaves and branches to improve the performance of the decision tree when it moves from the training data (where the classification is known) to real-world applications (where the classification is unknown – it is what you are trying to predict). The tree-building algorithm makes the best split at the root node where there are the largest number of records and, hence, a lot of information. Each subsequent split has a smaller and less representative population with which to work. Towards the end, idiosyncrasies of training records at a particular node display patterns that are peculiar only to those records. These patterns can become meaningless and sometimes harmful for prediction if you try to extend rules based on them to larger populations.

For example, say the classification tree is trying to predict height and it comes to a node containing one tall person named X and several other shorter people. It can decrease diversity at that node by a new rule saying "people named X are tall" and thus classify the training data. In a wider universe this rule can become less than useless. (Note that, in practice, we do not include irrelevant fields like "name", this is just an illustration.)

Pruning methods solve this problem – they let the tree grow to maximum size, then remove smaller branches that fail to generalize.

Since the tree is grown from the training data set, when it has reached full structure it usually suffers from over-fitting (i.e. it is "explaining" random elements of the training data that are not likely to be features of the larger population of data). This results in poor performance on real life data. Therefore, it has to be pruned using the validation data set .