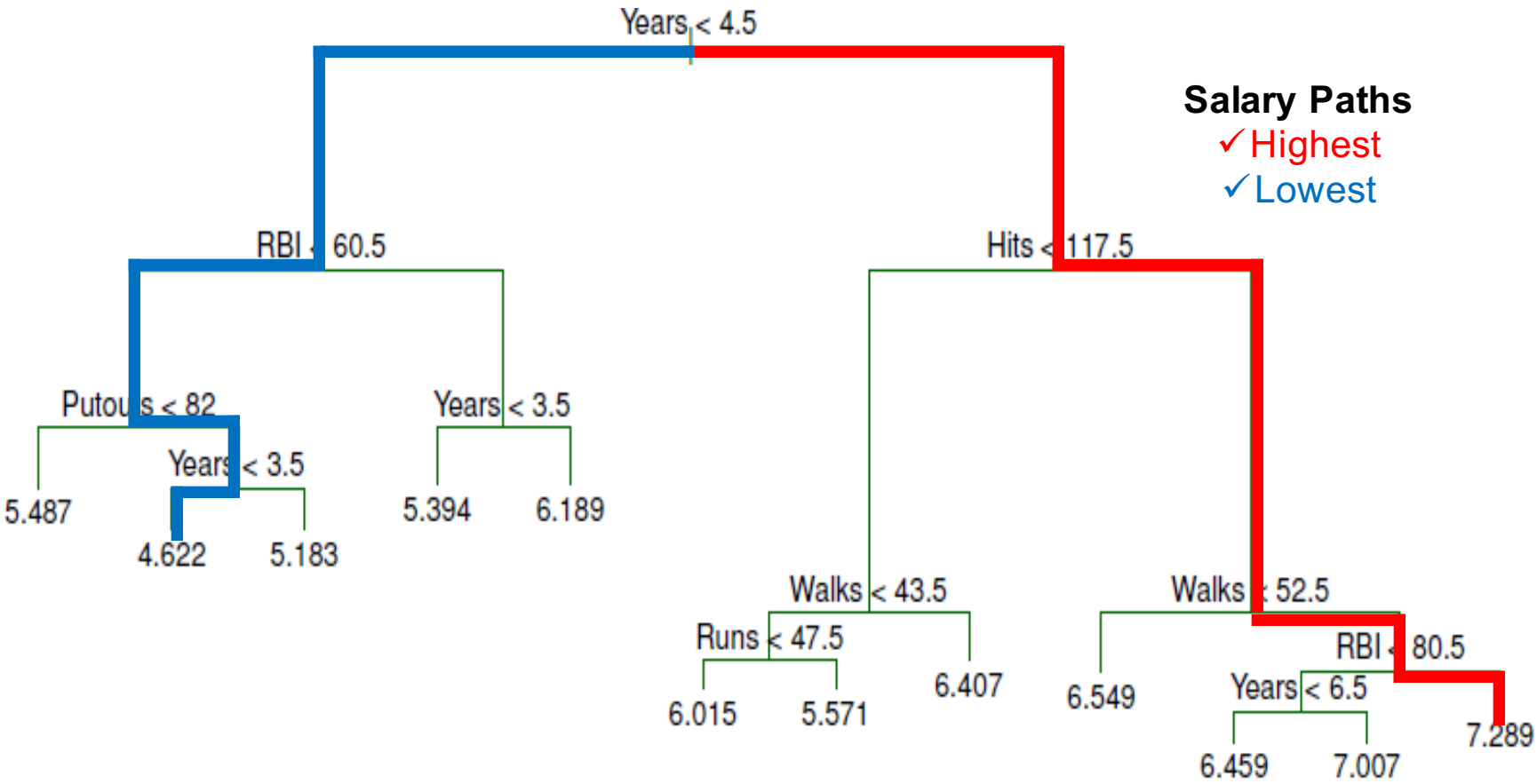
Growing Trees

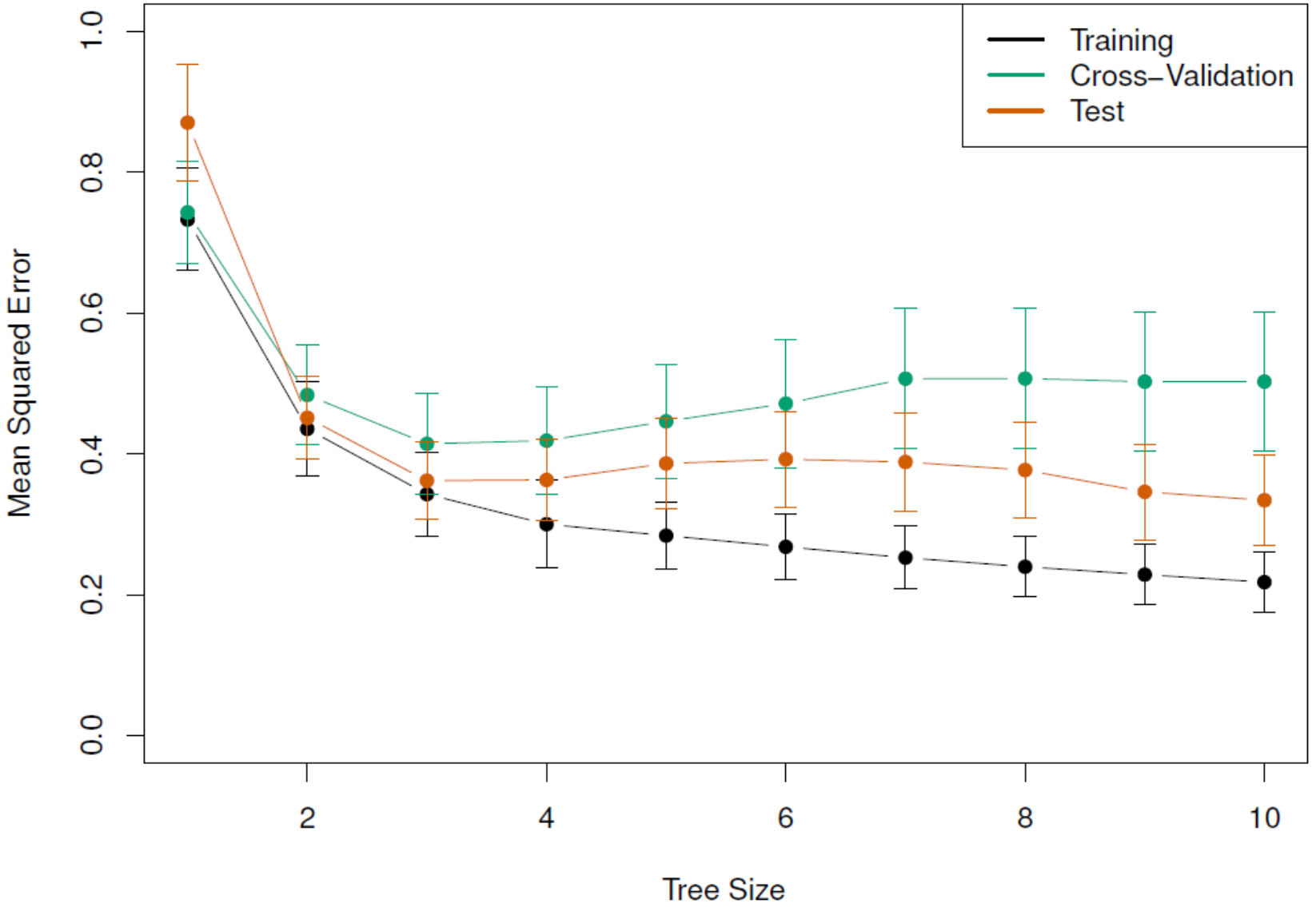
Where to Stop & Terminate Nodes

- Where do we **stop** the branching or segmenting?
- **Too few** branches will lead to **high bias** and **low accuracy**
- **Too many** branches will lead to very **low training** ESS.
- When the number of **regions** = the number of data **points**, we have **"Terminal Nodes"** and the training $ESS = 0$ (the nodes in between are called **"Internal Nodes"**)
- Thus, **too many** branches leads to **high variance** and **over-fitting**
- So, the **test ESS** will not necessarily be minimized and will most likely **bottom out** at some point
- Consequently, an important **goal** in regression trees is to find the **optimal** number of **branches**
- The general process is to **build** a relatively **large** regression **tree** with training data and then **"prune"** it back to a smaller **"sub-tree"** with a number of nodes/branches that **minimize** the **cross-validation** ESS

Baseball Salaries Illustration



Pruning Illustration



KOGOD SCHOOL
of
BUSINESS