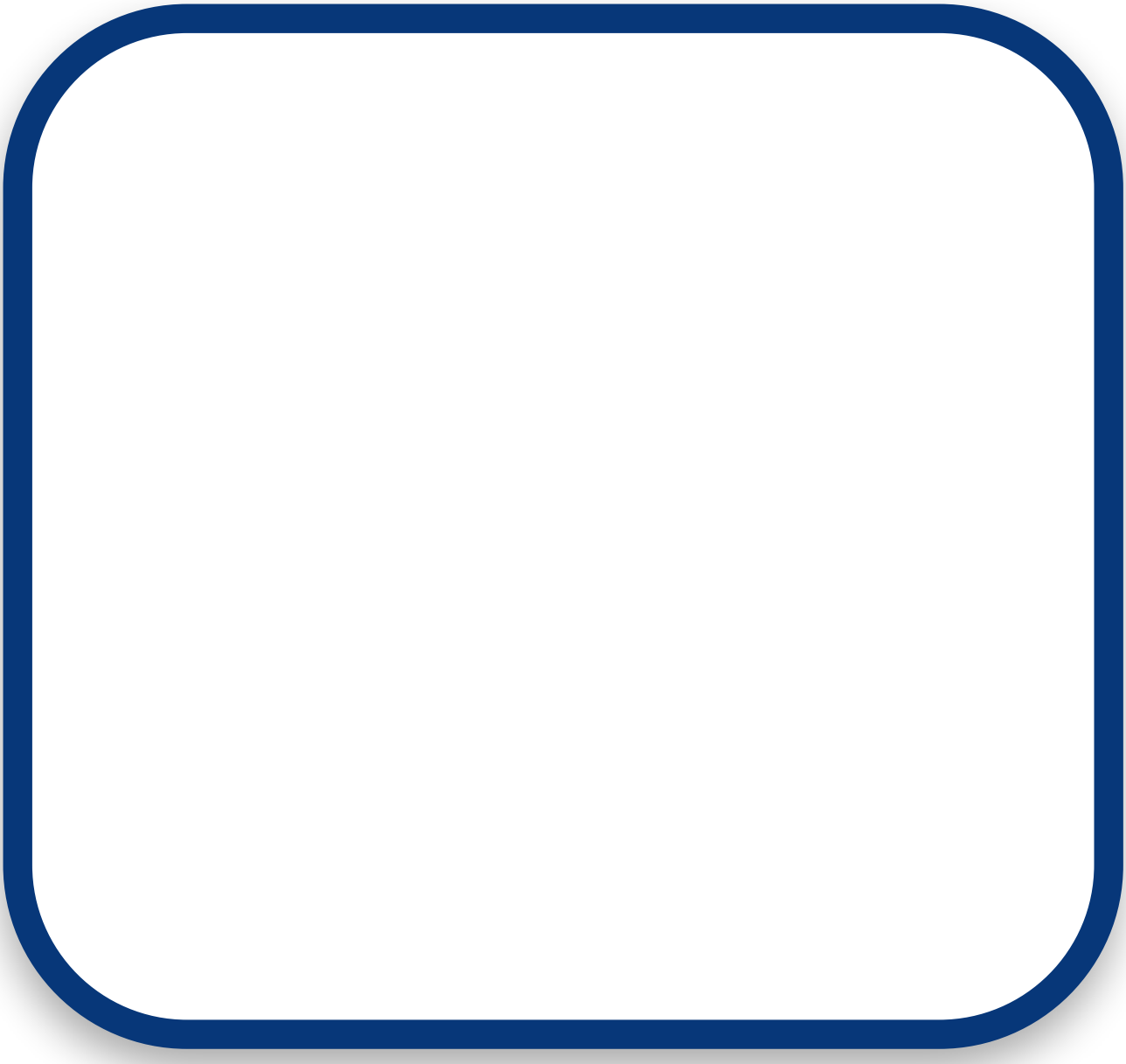
H2O.ai



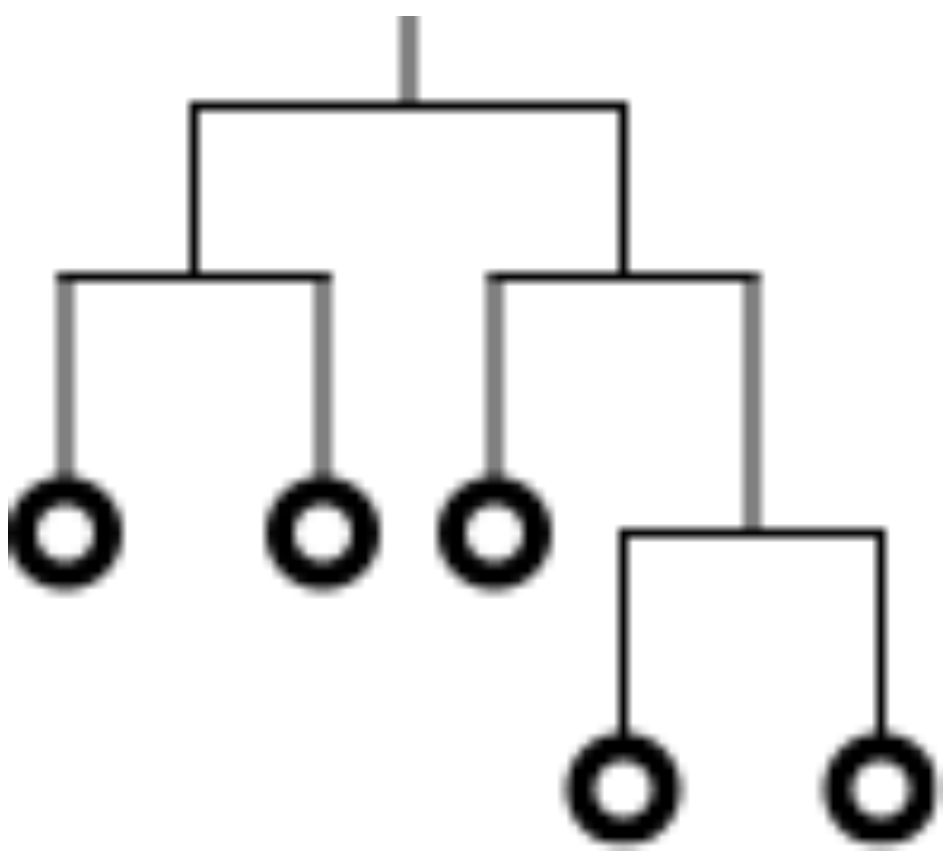
Distributed RFs

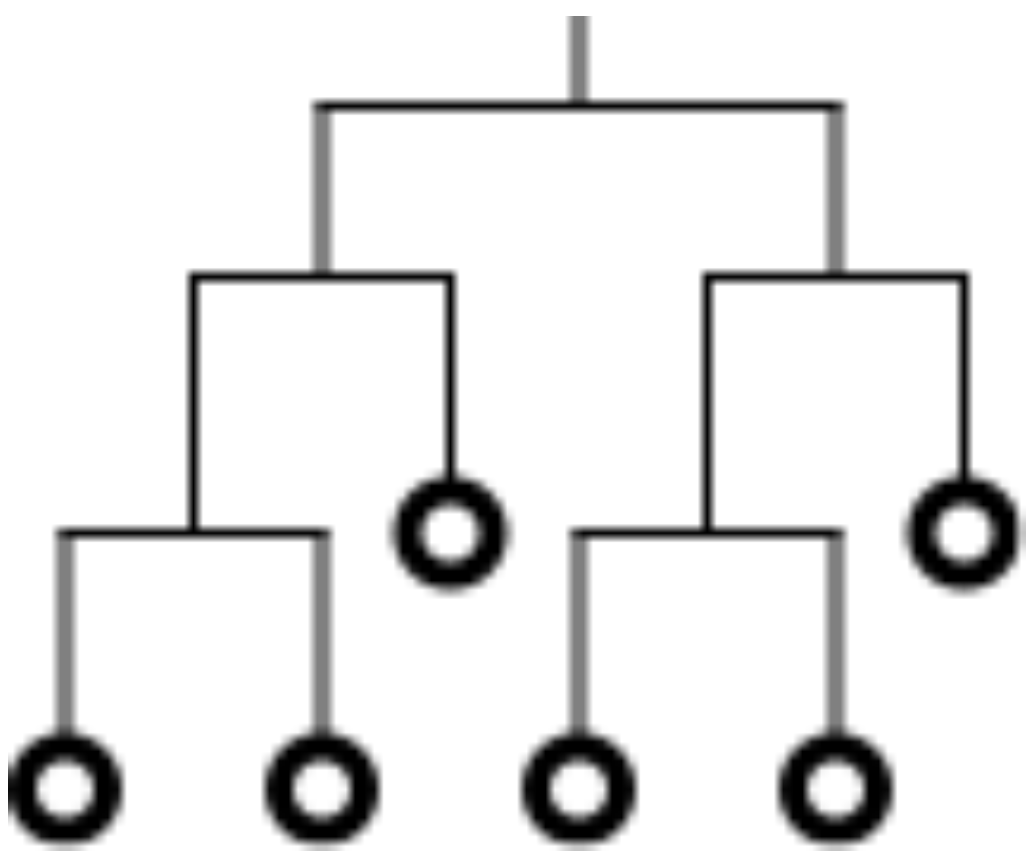
Method 1: Parallelize by tree

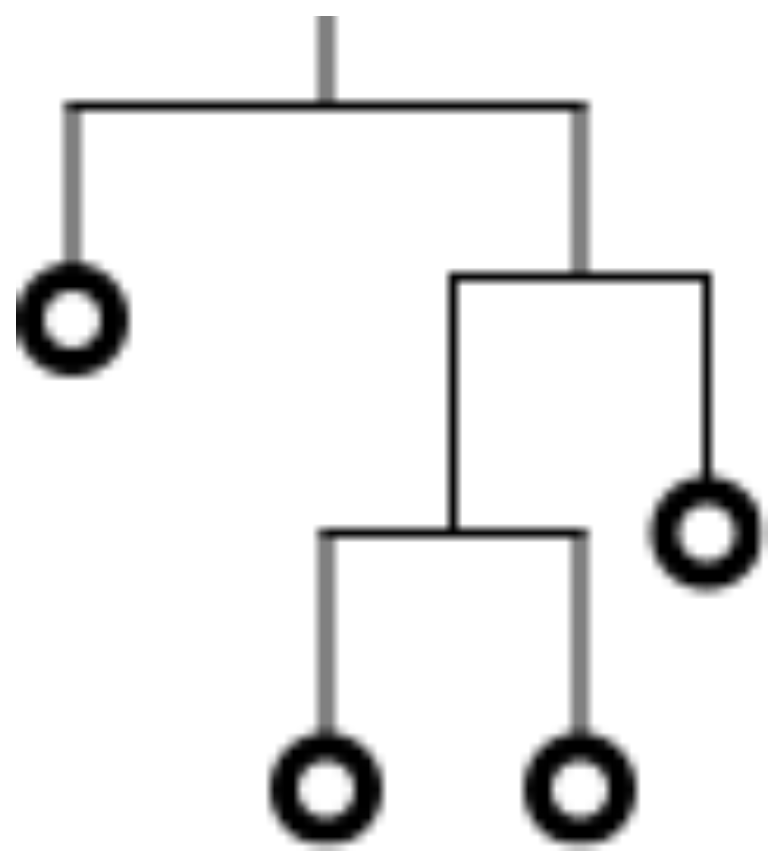
✔ low communication between nodes (only to reduce as part of map/reduce)

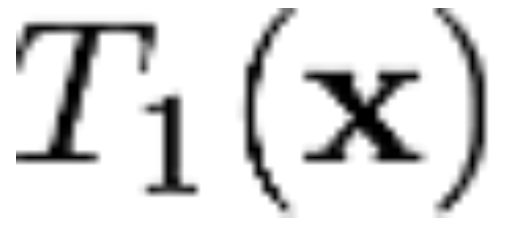
✗ every node must have all data in memory

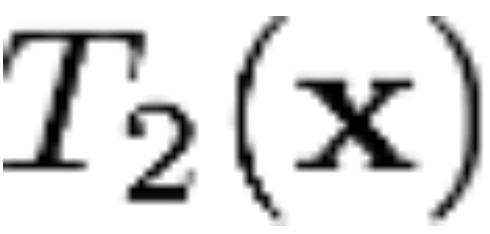
✗ does not generalize to GBMs

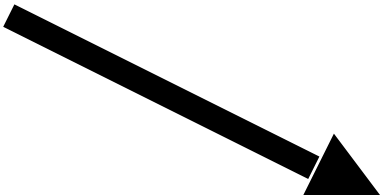


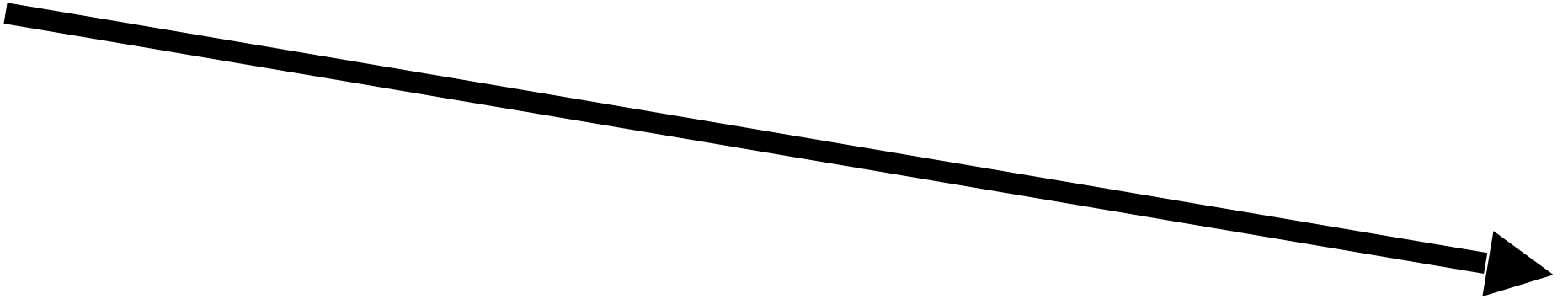


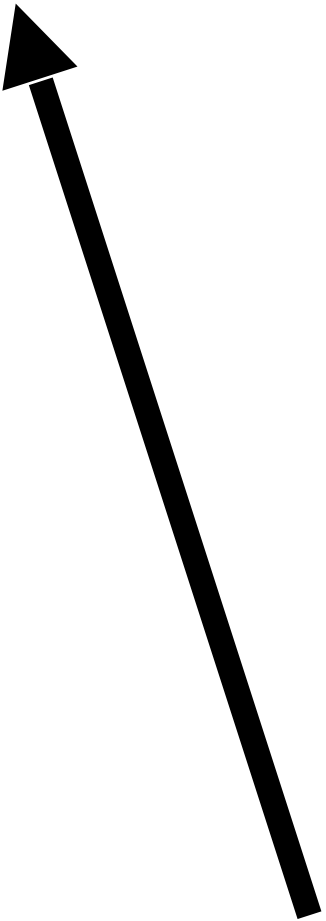


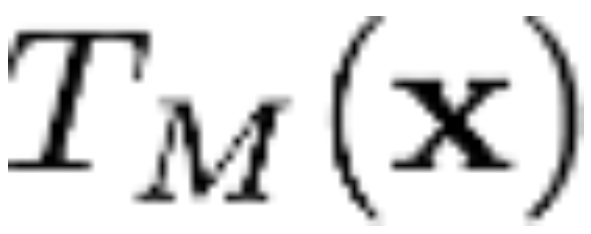








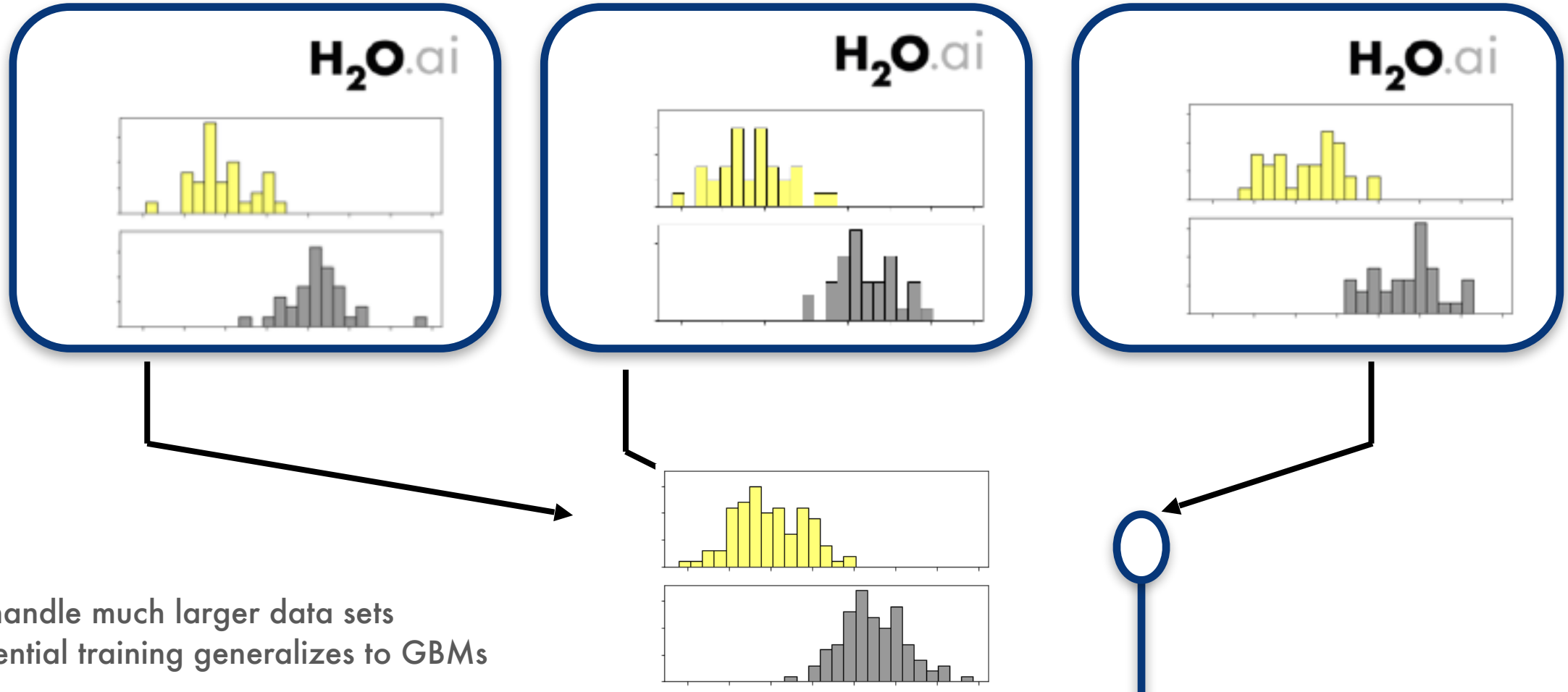




$$f_M(\mathbf{x}) = \sum_{i=1}^{M} T_i(\mathbf{x})$$

Distributed RFs in H2O

Method 2: Parallelize by data



✔ can handle much larger data sets
✔ sequential training generalizes to GBMs

✗ more communication between nodes to reduce the computations
✗ relies on approximate split calculation via histogram

H2O.ai

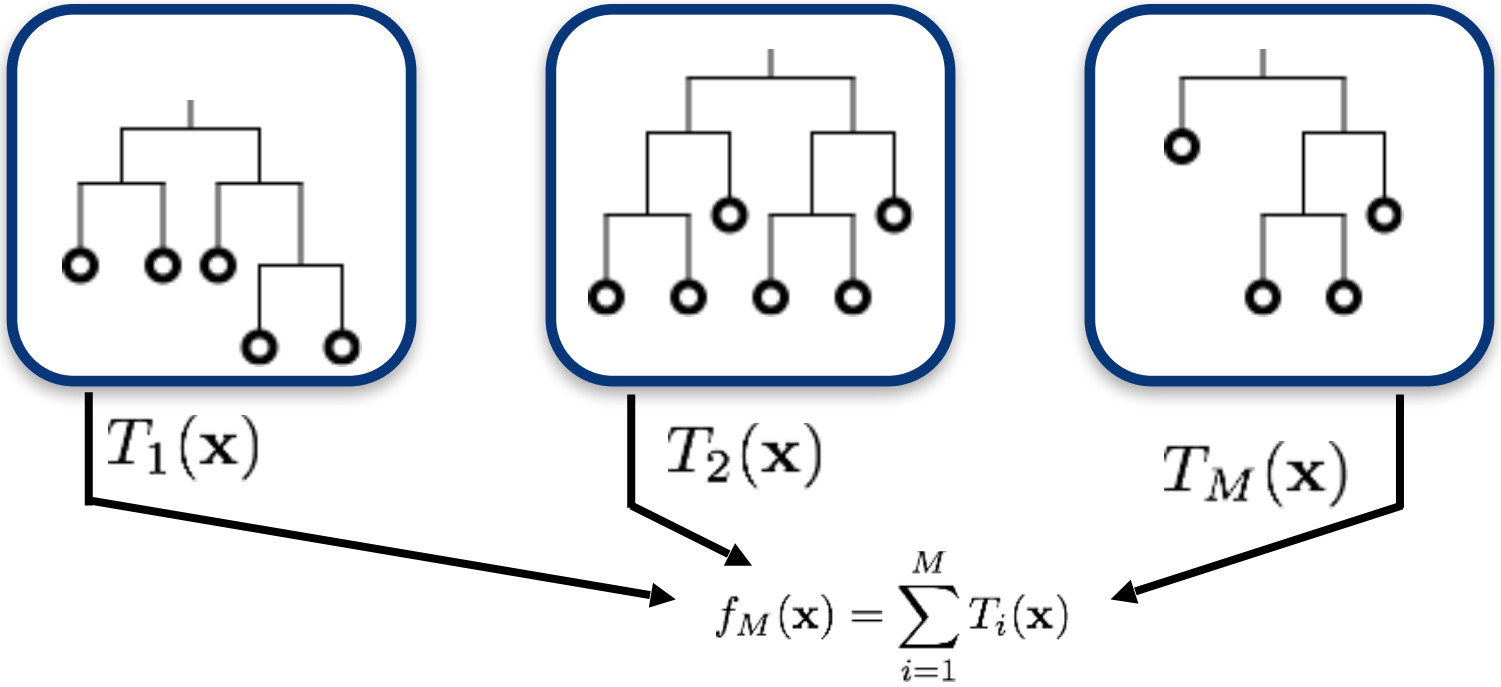
Distributed RFs

Method 1: Parallelize by tree

✔ low communication between nodes (only to reduce as part of map/reduce)

✗ every node must have all data in memory

✗ does not generalize to GBMs



H2O.ai