Ensemble method (1)
Random Forest

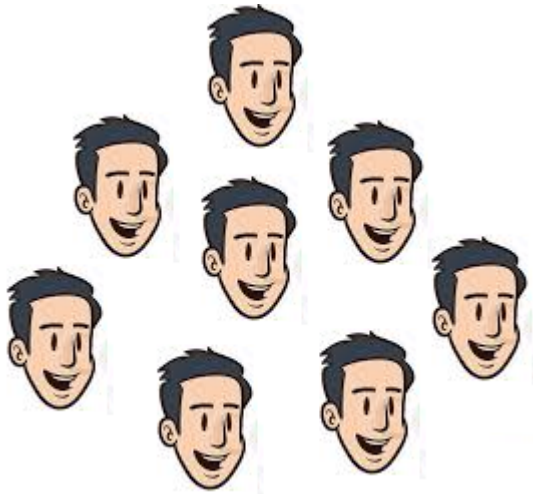
Ensemble

What is an Ensemble?

What is an Ensemble?

- An individual model might be a weak-learner,
- Aggregated models can predict better

Diversity matters

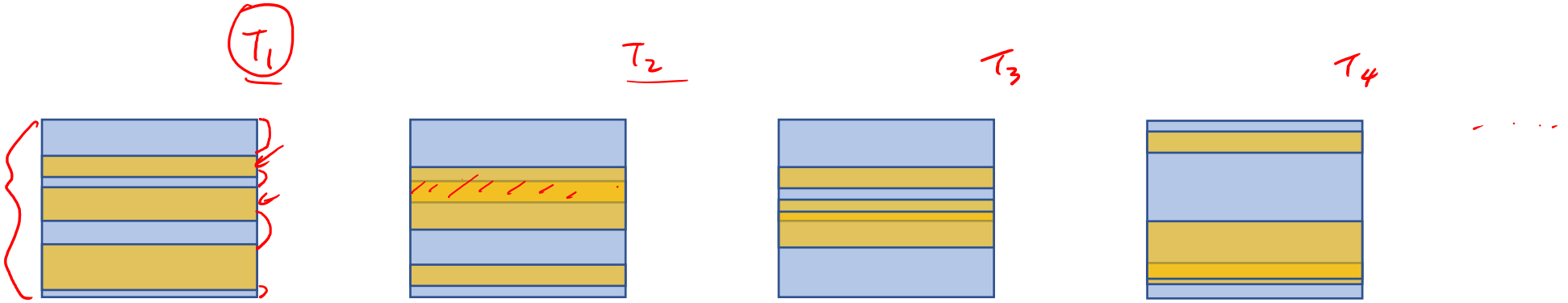


How do we diversify our models?

- Idea1: Models trained on different data subset

Bagging

Bagging (Bootstrap-Aggregation)



STEP1: Randomly sample a subset of training data with replacement (Bootstrap)

STEP2: Grow a tree (without pruning) on the subset of data

STEP3: Ensemble the result (regression : average, classification : vote)

Out of Bag error (OOB) : test the grown tree on the rest of data, then average

Random Forest

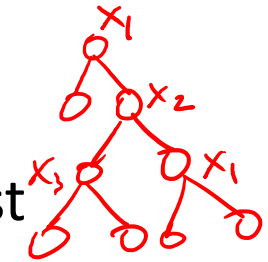
Bagging : random sampling of data

+

Decorrelation : random sampling of features

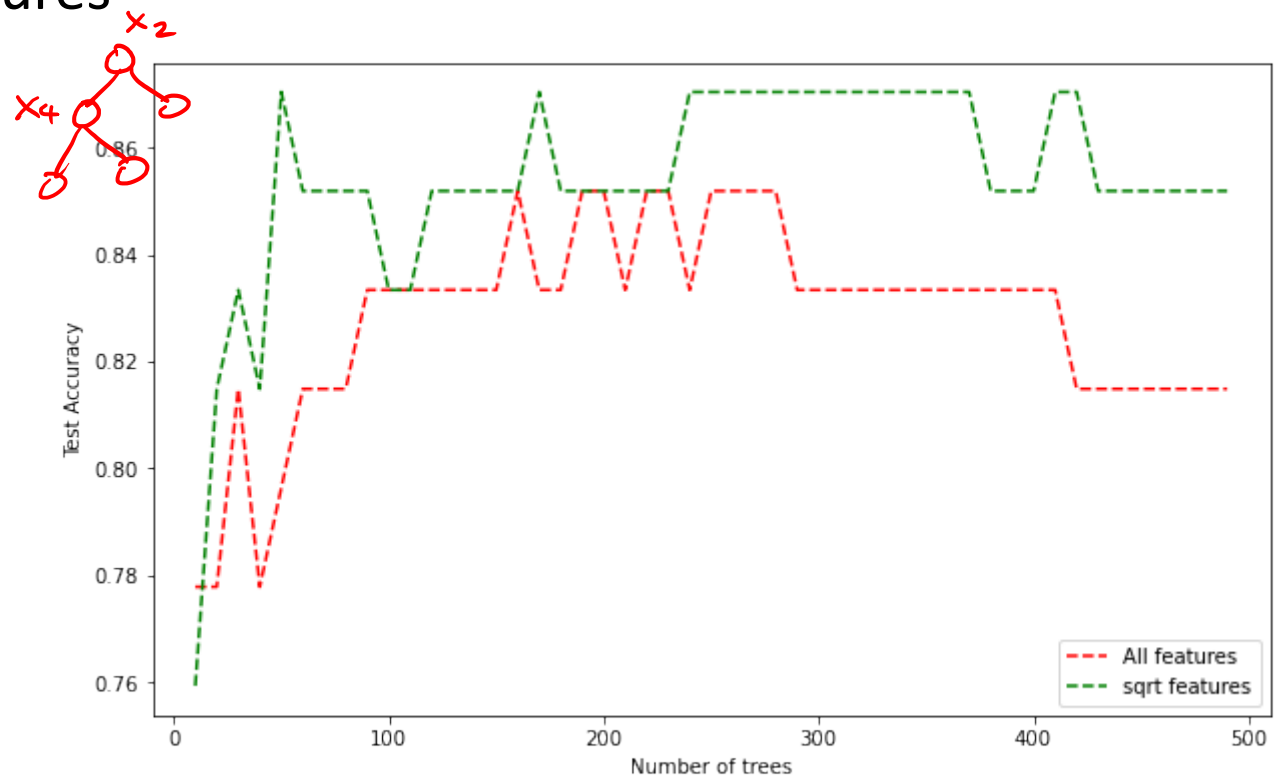
=

Random Forest

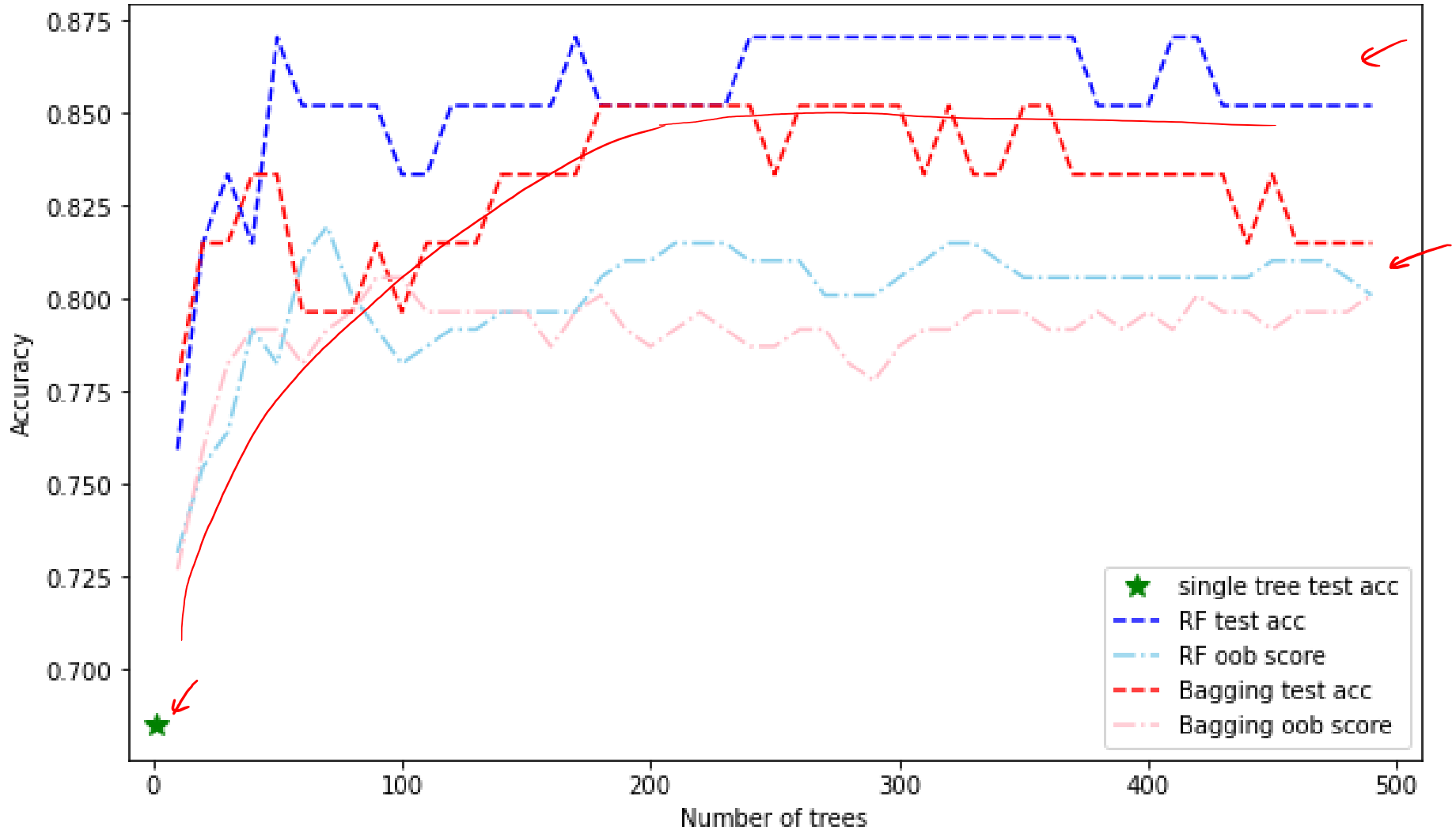


How do we sample features?

-> Rule of thumb : $\sqrt{n}$



Power of an ensemble of trees



Built-in feature importance