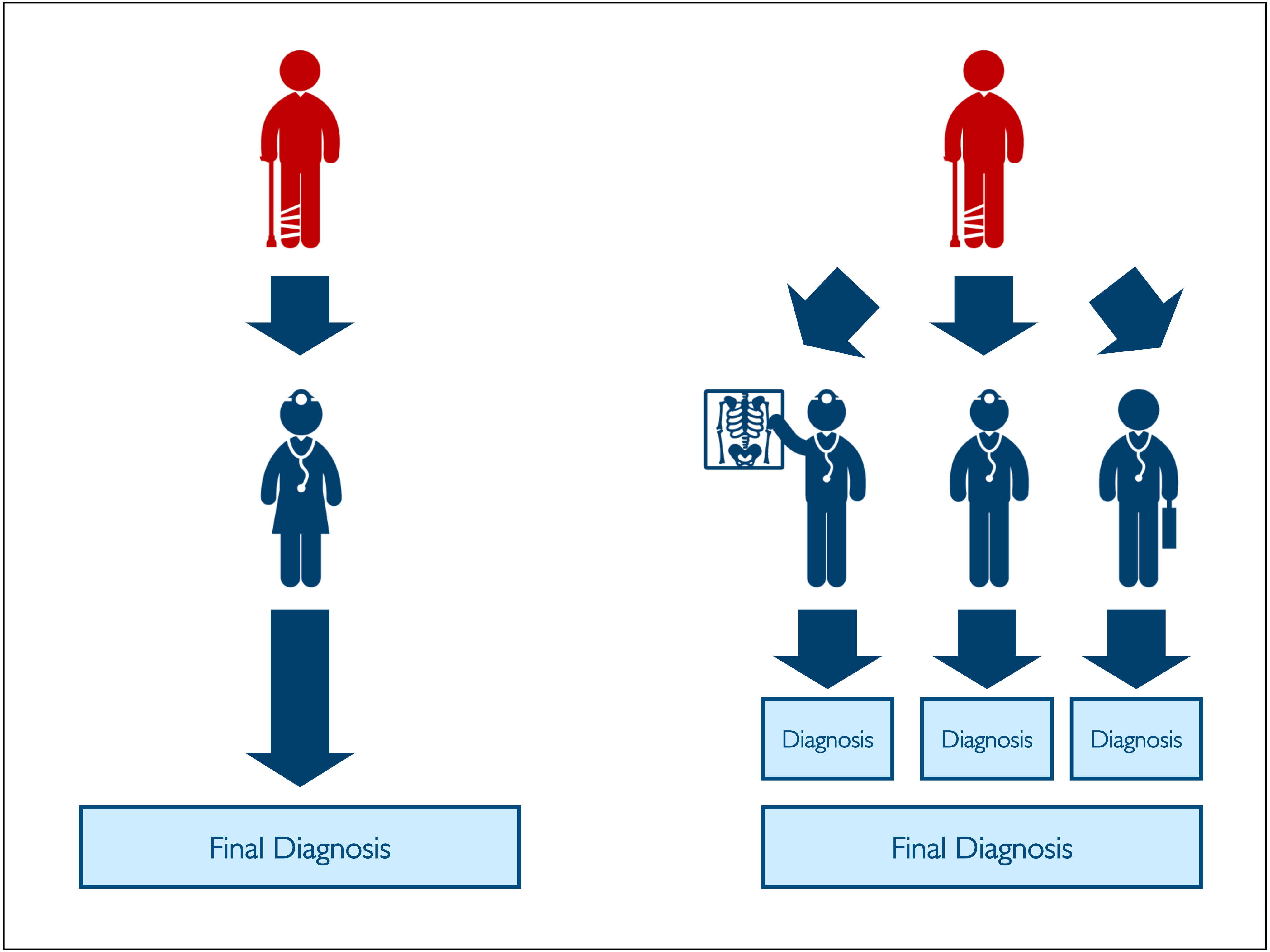
POLITECNICO DI MILANO

Classification Ensembles

Master in Analytics and Business Intelligence – Machine Learning

What are Ensemble Methods?


Diagnosis
Diagnosis
Diagnosis
Final Diagnosis
Final Diagnosis

Ensemble Methods

Generate a set of classifiers from the training data

Predict class label of previously unseen cases by aggregating predictions made by multiple classifiers

Building models ensembles

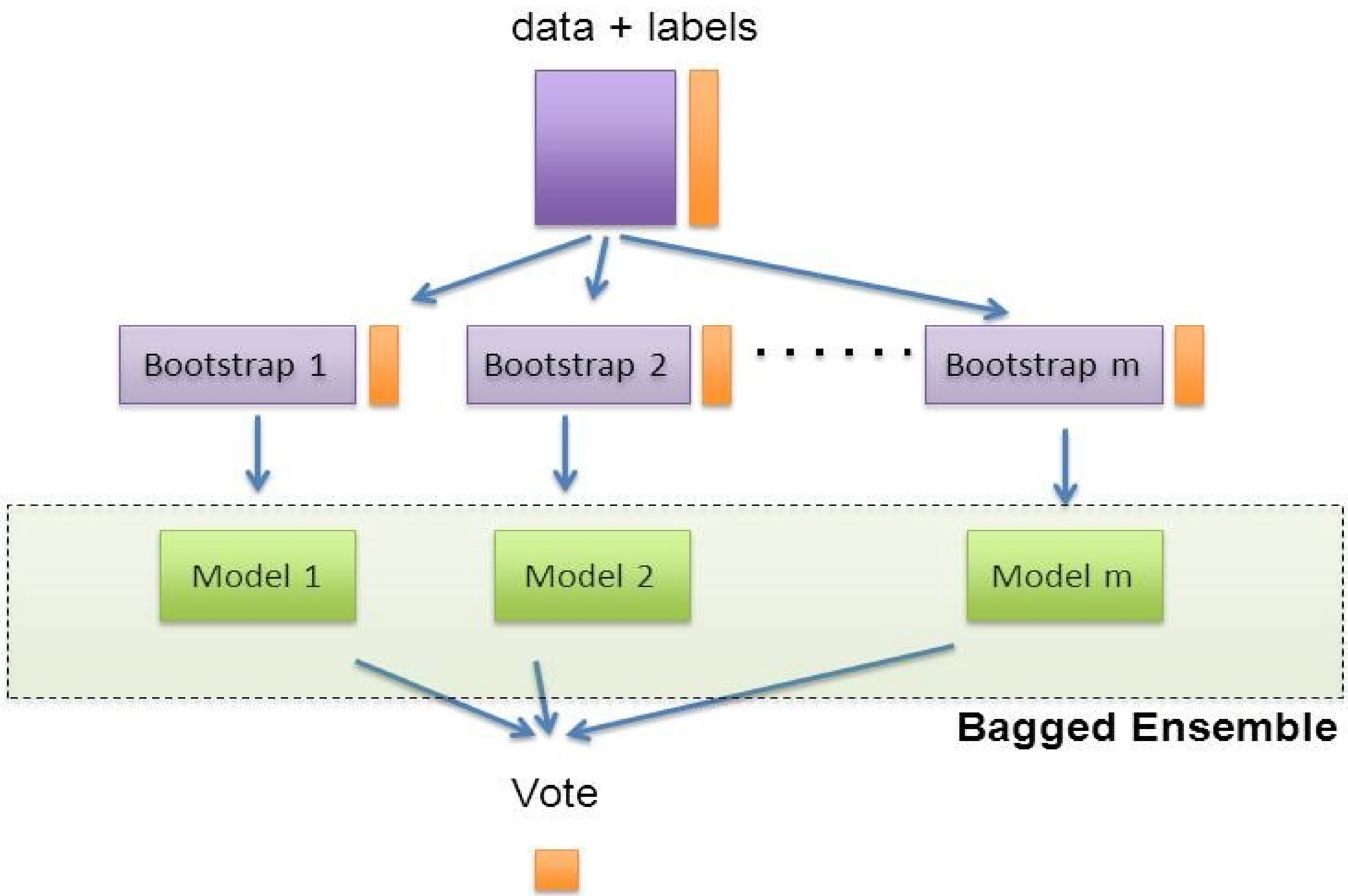
- Basic idea
  - Build different “experts”, let them vote
- Advantage
  - Often improves predictive performance
- Disadvantage
  - Usually produces output that is very hard to analyze

However, there are approaches that aim to produce a single comprehensible structure

how can we generate several models using
the same data and the same approach (e.g., Trees)?

Bagging

What is Bagging? (Bootstrap Aggregation)



Bagging works because it reduces variance by voting/averaging

usually, the more classifiers the better

it can help a lot if data are noisy

however, in some pathological hypothetical situations the overall error might increase

When Does Bagging Work?
Stable vs Unstable Classifiers

- A learning algorithm is unstable, if small changes to the training set cause large changes in the learned classifier
- If the learning algorithm is unstable, then bagging almost always improves performance
- Bagging stable classifiers is not a good idea
- Decision trees, regression trees, linear regression, neural networks are examples of unstable classifiers
- K-nearest neighbors is a stable classifier

Random Forests

the more uncorrelated the trees,

the better the variance reduction

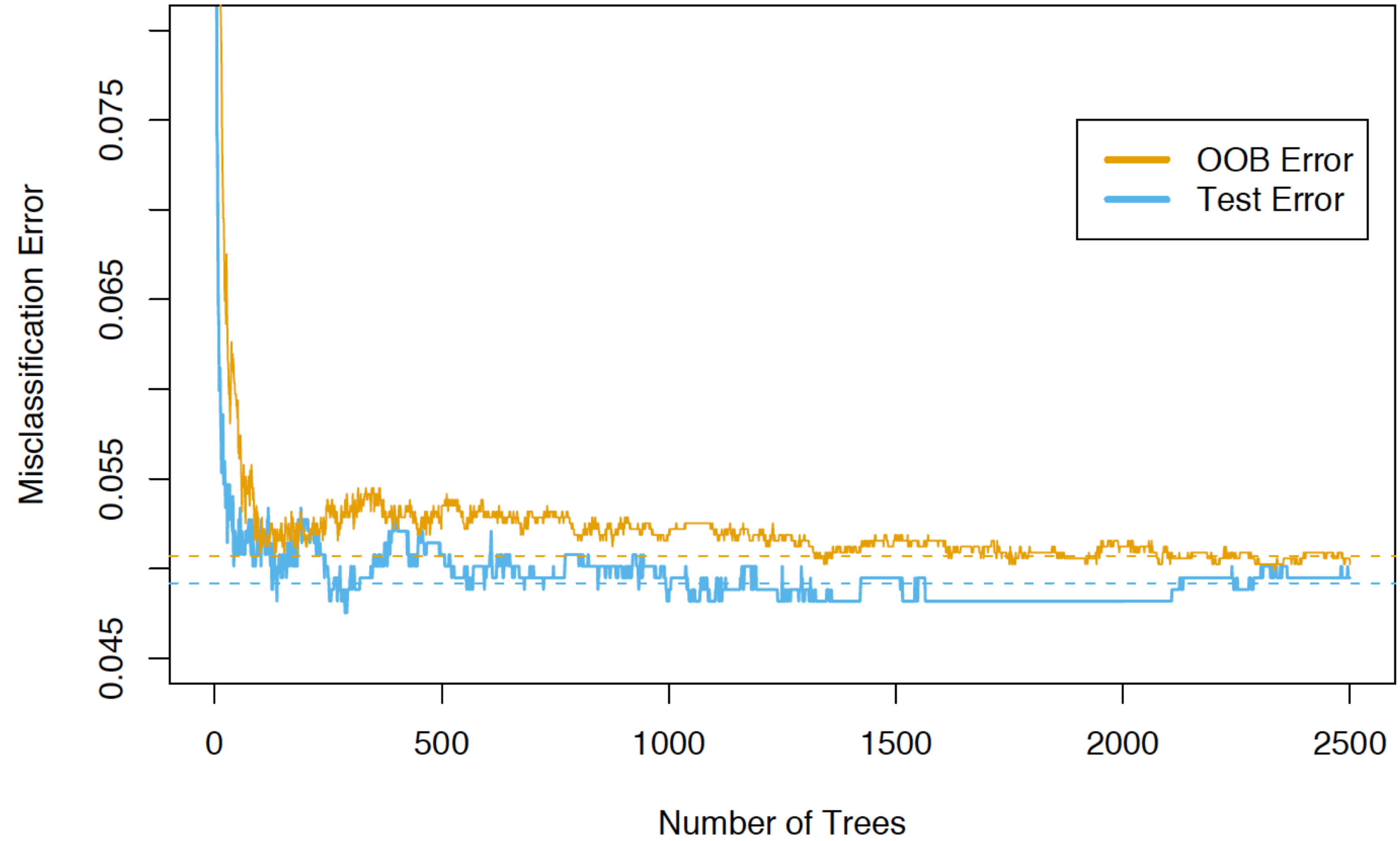
learning ensemble consisting of a bagging of unpruned decision tree learners with randomized selection of features at each split

What is a Random Forest?

- Random forests (RF) are a combination of tree predictors
- Each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest
- The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them
- Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, and are more robust with respect to noise

Out of Bag Samples

- For each observation (x_i, y_i), construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which the observation did not appear
- The OOB error estimate is almost identical to that obtained by n-fold crossvalidation and related to the leave-one-out evaluation
- Thus, random forests can be fit in one sequence, with cross-validation being performed along the way
- Once the OOB error stabilizes, the training can be terminated


OOB Error
Test Error
Misclassification Error
0.045
0.055
0.065
0.075
0
500
1000
1500
2000
2500
Number of Trees

Properties of Random Forests

- Easy to use ("off-the-shelve"), only 2 parameters (no. of trees, %variables for split)
- Very high accuracy
- No overfitting if selecting large number of trees (choose high)
- Insensitive to choice of split% (~20%)
- Returns an estimate of variable importance
- Random forests are an effective tool in prediction.
- Forests give results competitive with boosting and adaptive bagging, yet do not progressively change the training set.
- Random inputs and random features produce good results in classification - less so in regression.
- For larger data sets, we can gain accuracy by combining random features with boosting.