CS 4104
APPLIED MACHINE LEARNING

Dr. Hashim Yasin

National University of Computer and Emerging Sciences,

Faisalabad, Pakistan.

BOOSTING & BAGGING

BOOSTING & BAGGING

Boosting & Bagging

- Two strategies have been developed for producing optimal trees

Boosting:

- develop ***new classification trees*** based on the results of previous classification trees

Bagging:

- uses ***subsets of the training data*** to develop new classification trees

BAGGING

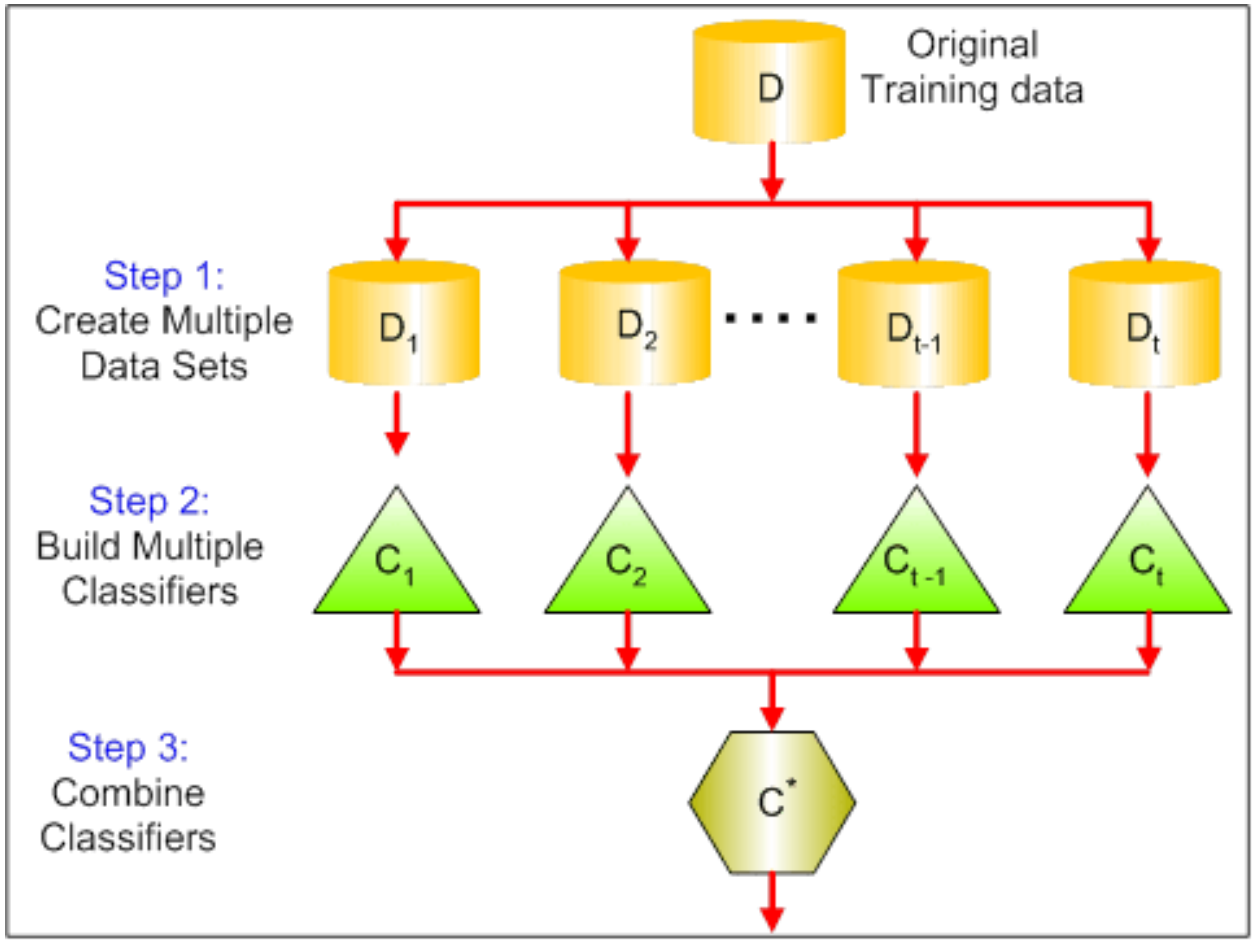
Bagging

- If *we split the data in random different ways*, decision trees give different results, results into **high variance**.
- We want to reduce the variance of a decision tree.

Bagging:

- **B**ootstrap **agg**regat**ing** is a method that result in *low variance*.

Bagging


D
Original
Training data
Step 1:
Create Multiple
Data Sets
D1
D2
Dt-1
Dt
Step 2:
Build Multiple
Classifiers
C1
C2
Ct -1
Ct
Step 3:
Combine
Classifiers
C*

Bagging

- For each sample *b*, we calculate $f^b(x)$, then:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

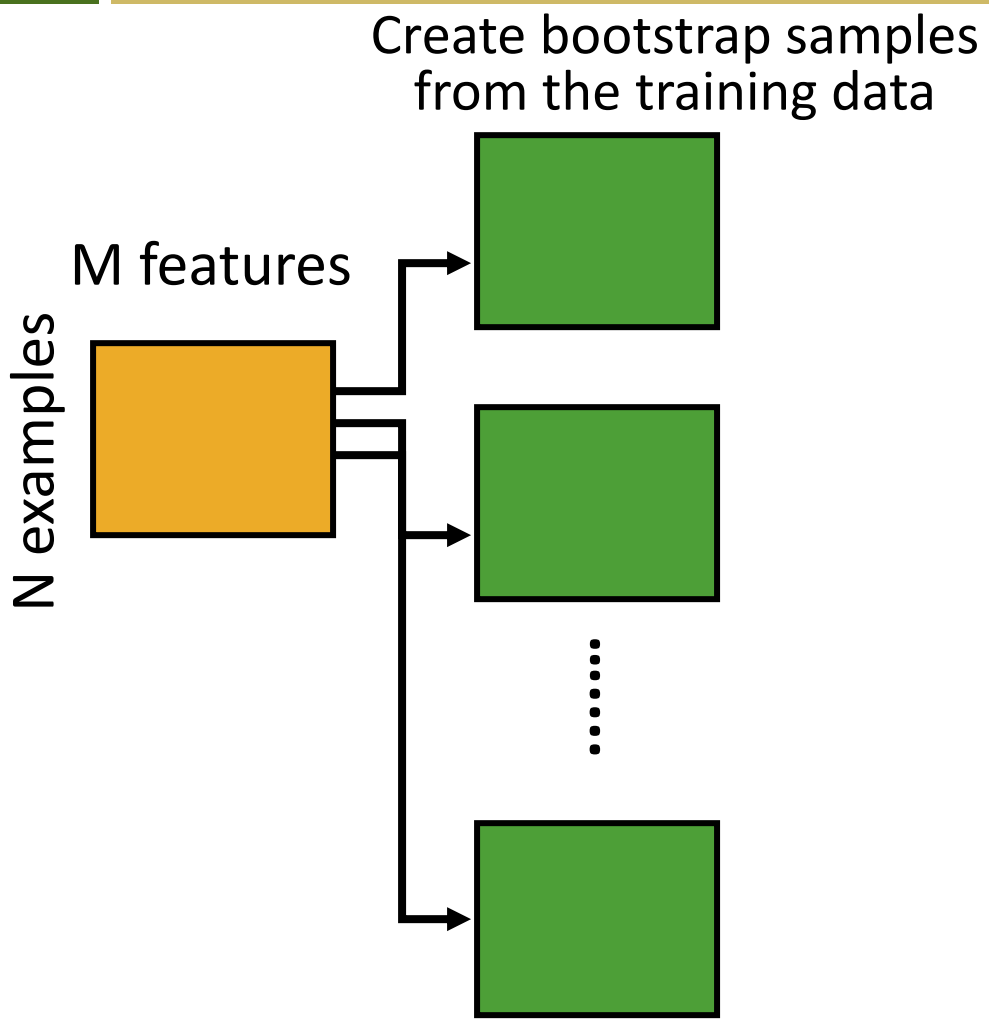
The prediction at input x when bootstrap sample *b* is used for training

How?

Bootstrap

- Construct B (hundreds) of trees (no pruning)
- Learn a classifier for each bootstrap sample and average them
- Very effective method

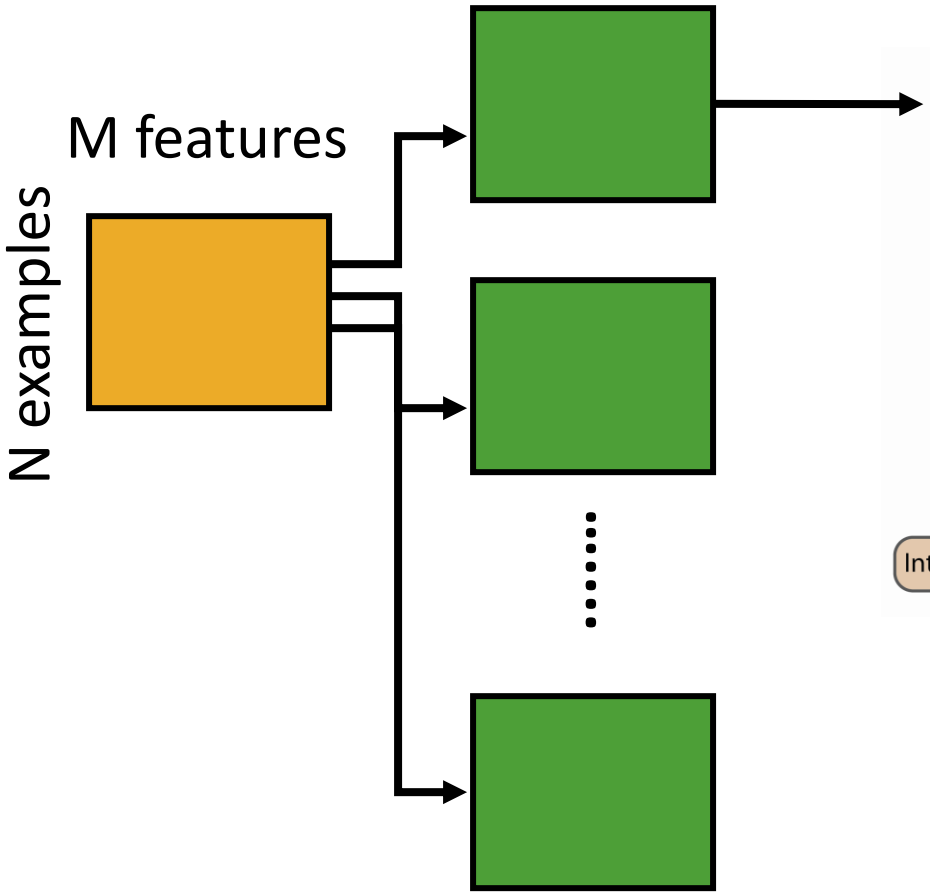
Bagging

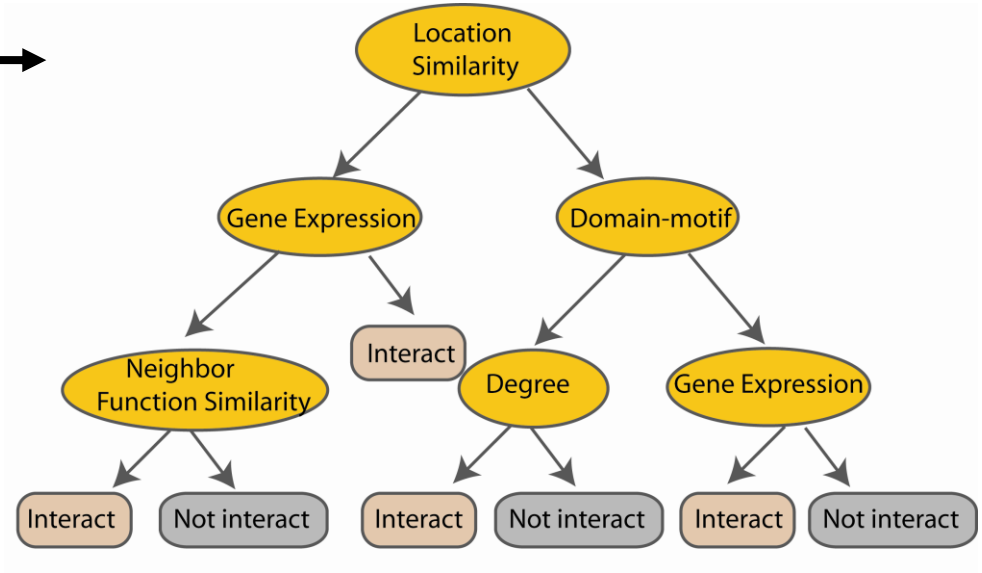

Create bootstrap samples
from the training data
M features
N examples

Bagging

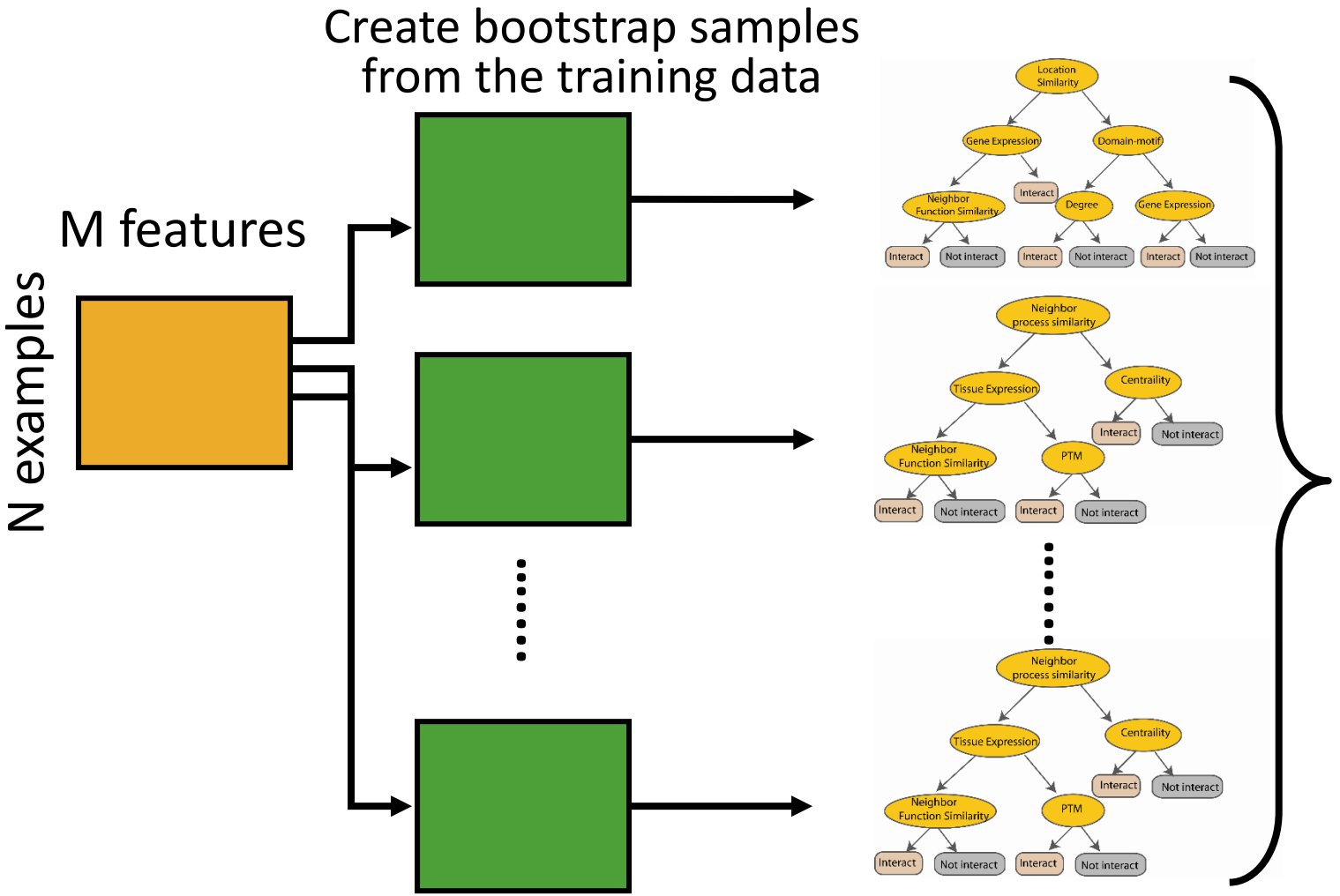
Create bootstrap samples from the training data

Construct a decision tree





Bagging


Create bootstrap samples from the training data
M features
N examples
Location Similarity
Gene Expression
Domain-motif
Neighbor Function Similarity
Interact
Degree
Gene Expression
Interact
Not interact
Interact
Not interact
Interact
Not interact
Neighbor process similarity
Tissue Expression
Centrality
Interact
Not interact
Neighbor Function Similarity
PTM
Interact
Not interact
Interact
Not interact
Neighbor process similarity
Tissue Expression
Centrality
Interact
Not interact
Neighbor Function Similarity
PTM
Interact
Not interact
Interact
Not interact
Take the majority vote

Bagging

- Reduces overfitting (variance)
- Normally uses one type of classifier
  - Decision trees are popular
- Easy to parallelize
- Bagging results in improved accuracy over prediction using a single tree
- Unfortunately, difficult to interpret the resulting model.
  - Bagging improves prediction accuracy at the expense of interpretability.

Bagging … Issues

- Each tree is identically distributed
- the expectation of the average of B such trees is the same as the expectation of any one of them
- the bias of bagged trees is the same as that of the individual trees

Bagging generate correlated trees.

RANDOM FOREST

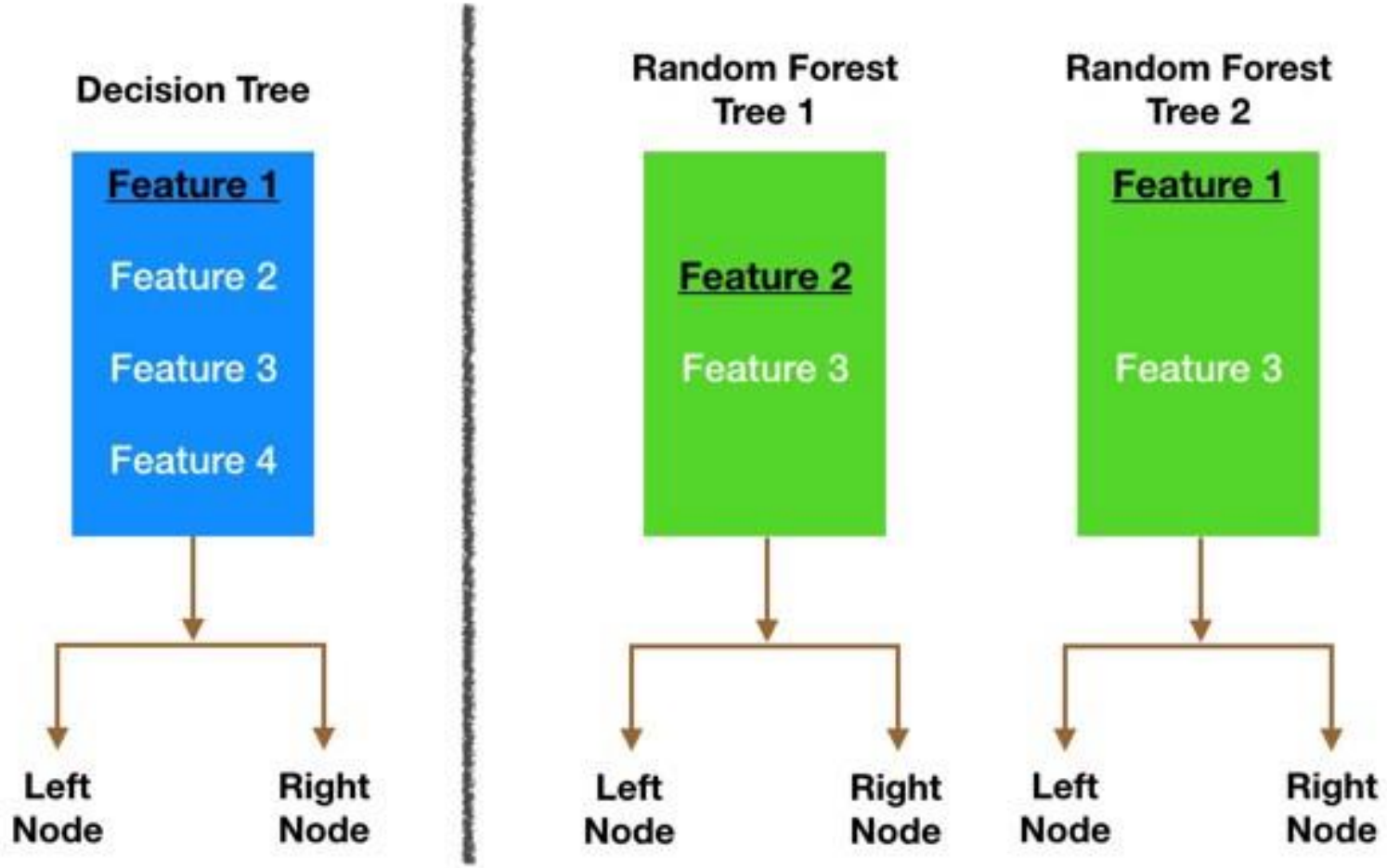
Random Forest

- **Random forest** is a **bagging** technique and **not a boosting** technique.
- The trees in **random forests** run in parallel.
- There is no interaction between these trees while building the trees.

Random Forest

- Given n observations with p predictors.
- Input:
  - $m << p$ the fraction of the predictors to sample at each split (often $m = \sqrt{p}$)
  - f, the fraction of the data to use for training
  - k, the number of trees in the forest.
- Repeat k times:
  - Choose a training set by choosing $f \times n$ training cases (with replacement). This is called bagging
  - Build a decision tree as follows
    - For each node of the tree, randomly choose m variables and find the best split from among those m variables
    - Repeat until the full tree is built (no pruning)

Random Forest



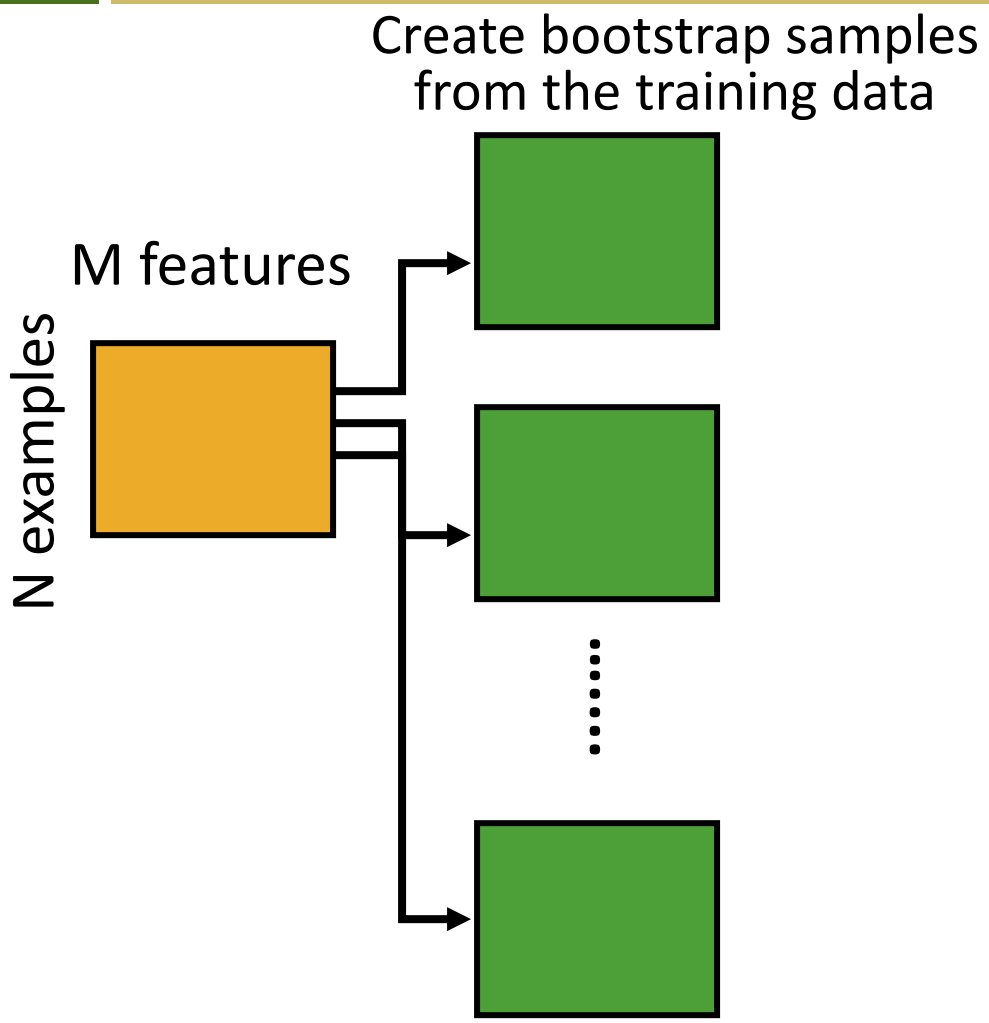
Random Forest

- To make a prediction at a new point *x*, we do:

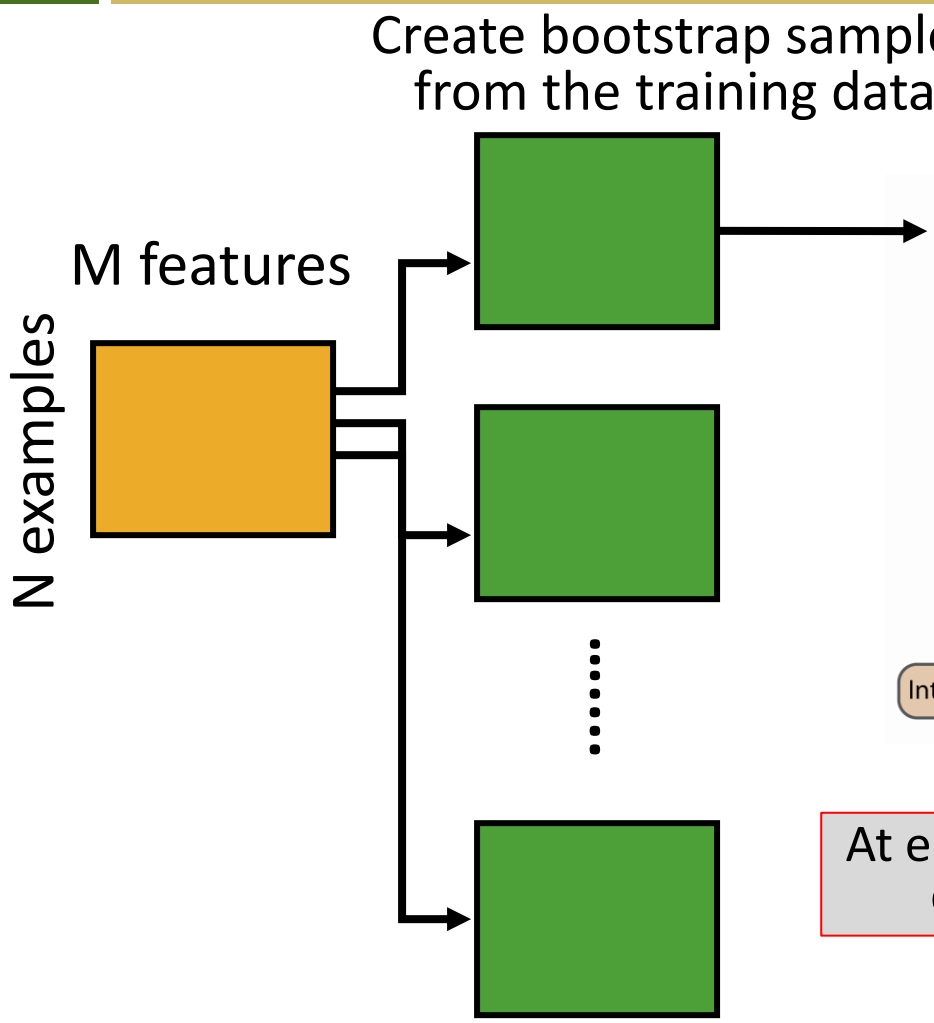
- **For regression:**
  - average the results

- **For classification:**
  - majority vote

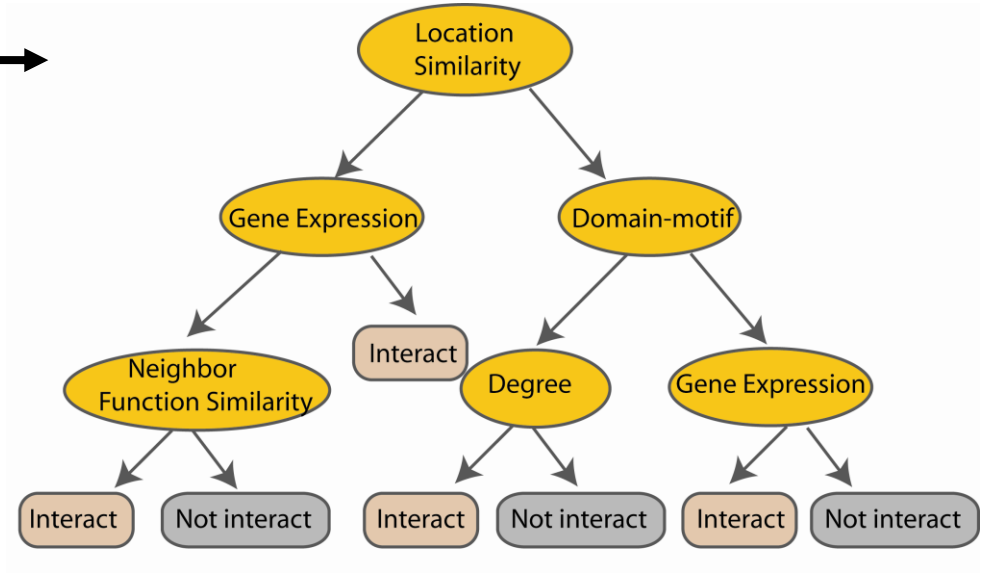
Random Forest



Random Forest

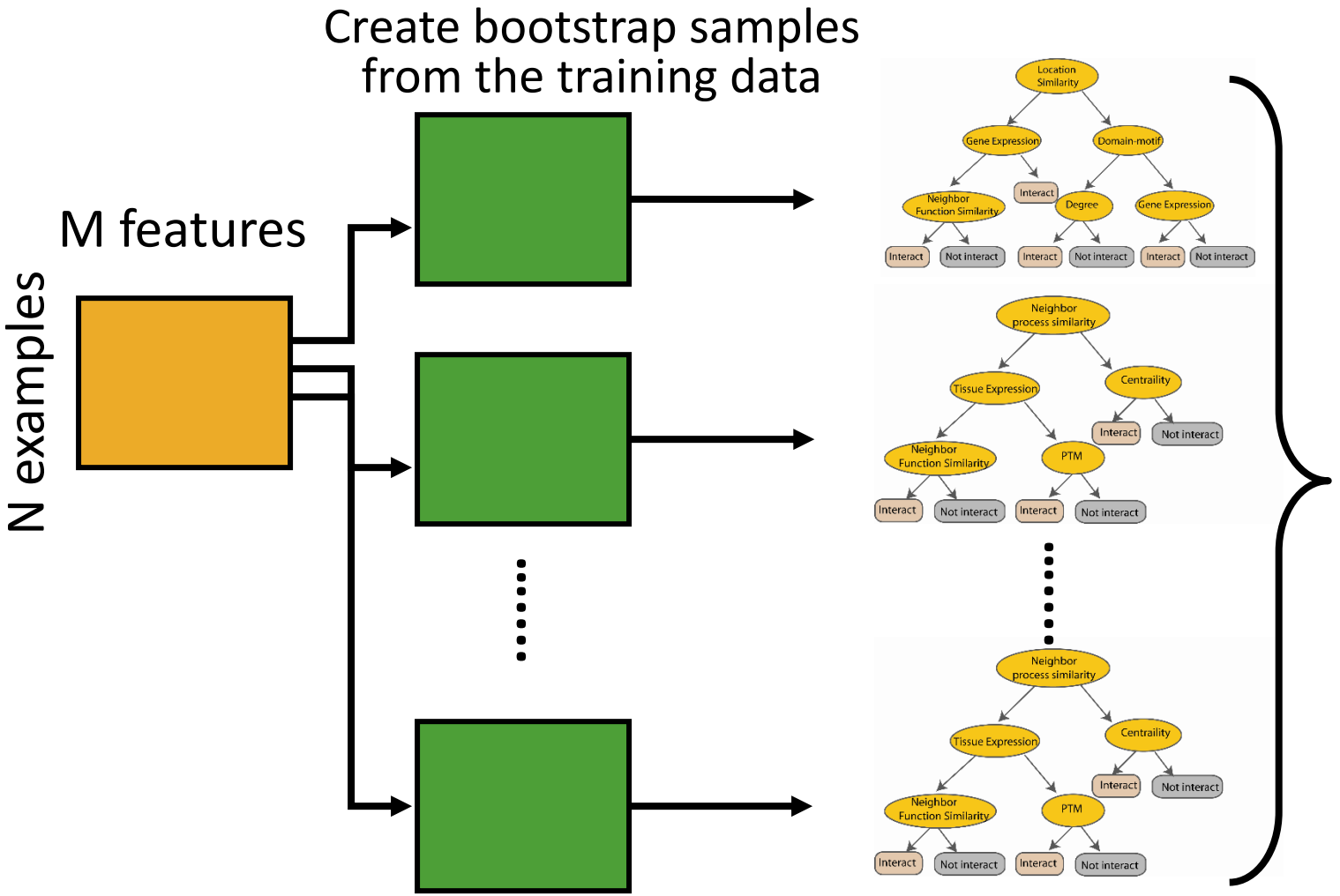


Construct a decision tree

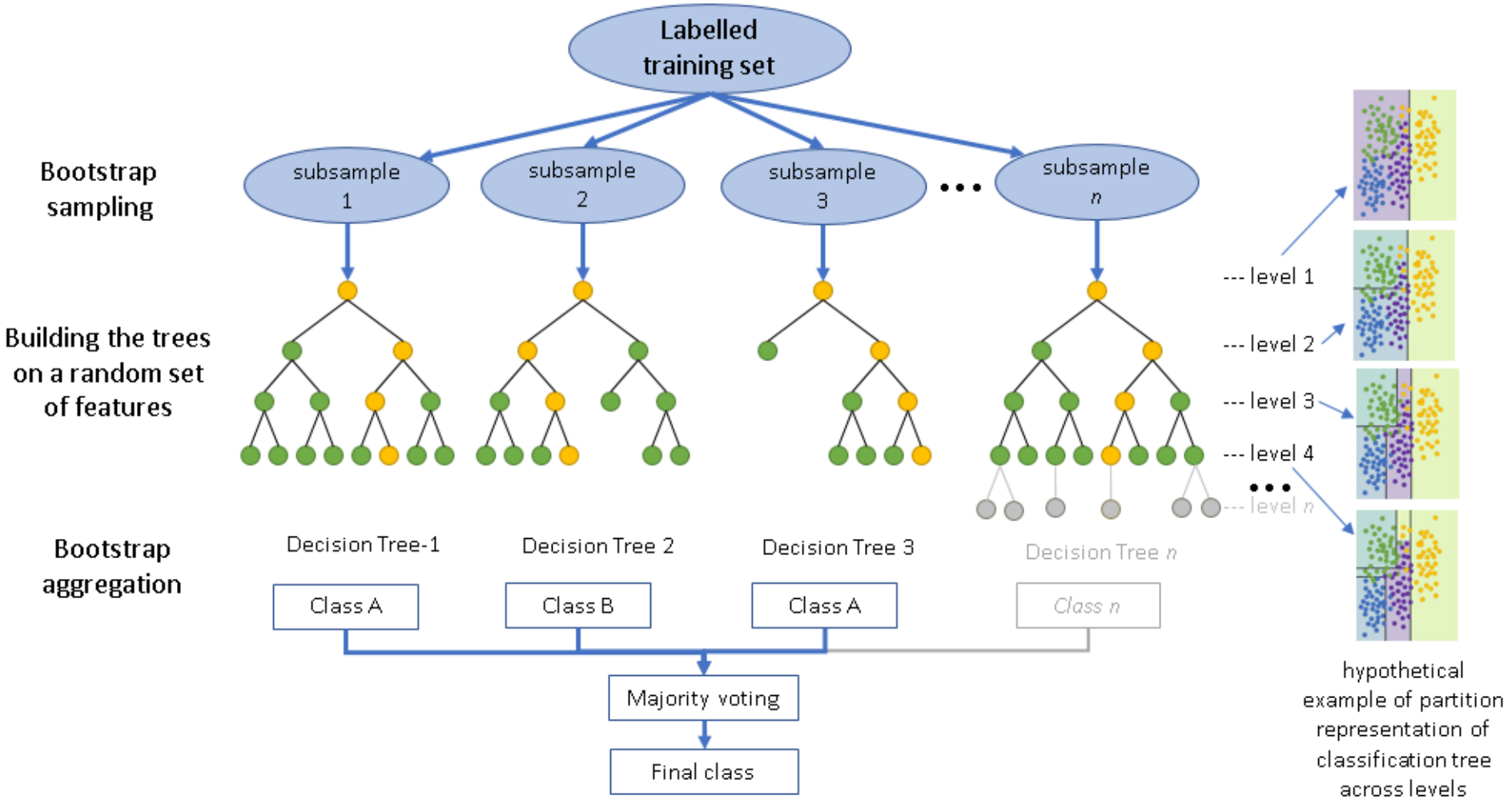


At each node in choosing the **split feature** choose only among $m<M$ features

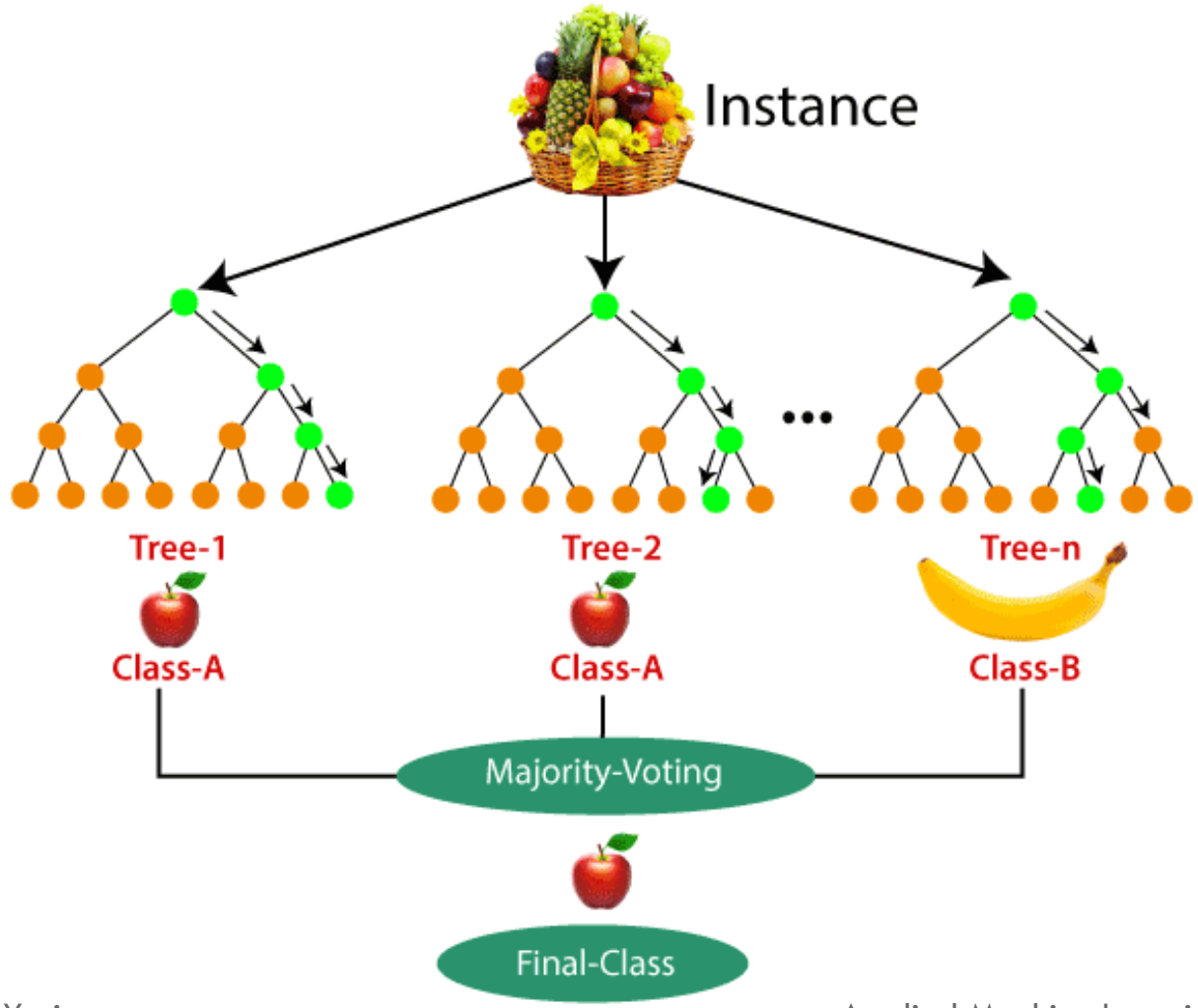
Random Forest



Random Forest



Random Forest



Random Forest

- It is one of the most accurate learning algorithms available.
  - For many datasets, it produces a **highly accurate classifier.**
- It runs efficiently on large databases.
- It can **handle thousands of input variables** without variable deletion.
- It gives estimates of what variables that are important in the classification.

Random Forest

- It generates an internal **unbiased estimate of the generalization error** as the forest building progresses.
- It has an **effective method for estimating missing data**,
  - maintains accuracy when a large proportion of the data are missing.

Random Forest … Issues

- Random forests have been observed to **overfit for some datasets** with noisy classification/regression tasks.

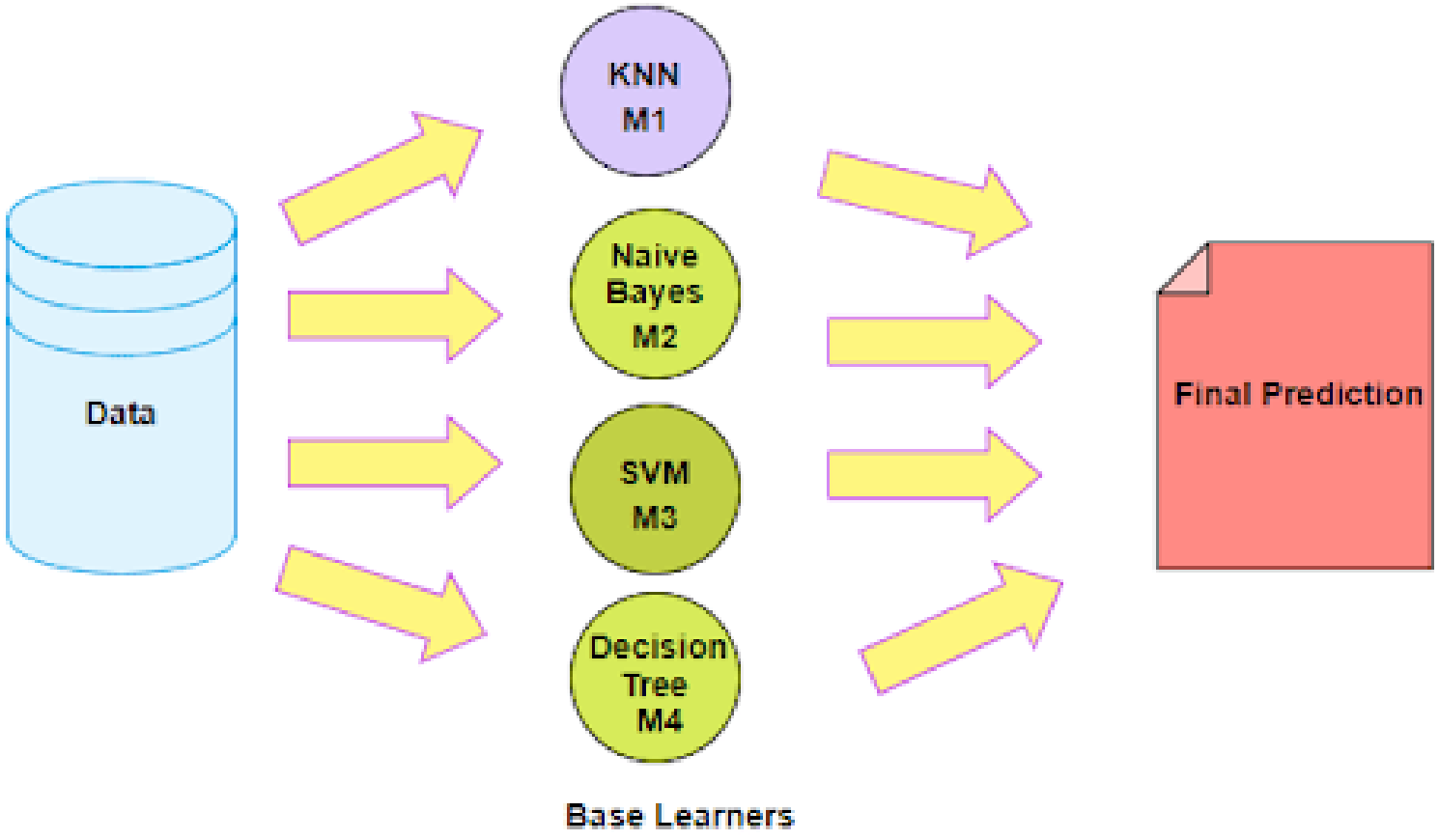
- For data including *categorical variables with different number of levels*, **random forests are biased in favor of those attributes with more levels.**

ENSEMBLE LEARNING

Ensemble Learning

- An Ensemble method is a technique that **combines**
  - **the predictions from multiple machine learning algorithms** together to make more accurate predictions than any individual model.
- A model comprised of many models is called an **Ensemble model.**

Ensemble Learning



Acknowledgement

Tom Mitchel, Russel & Norvig, Andrew Ng, Alpydin & Ch. Eick.