Ensembles

What are ensembles

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.

What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.
- Any predictor can be used as a base predictor.

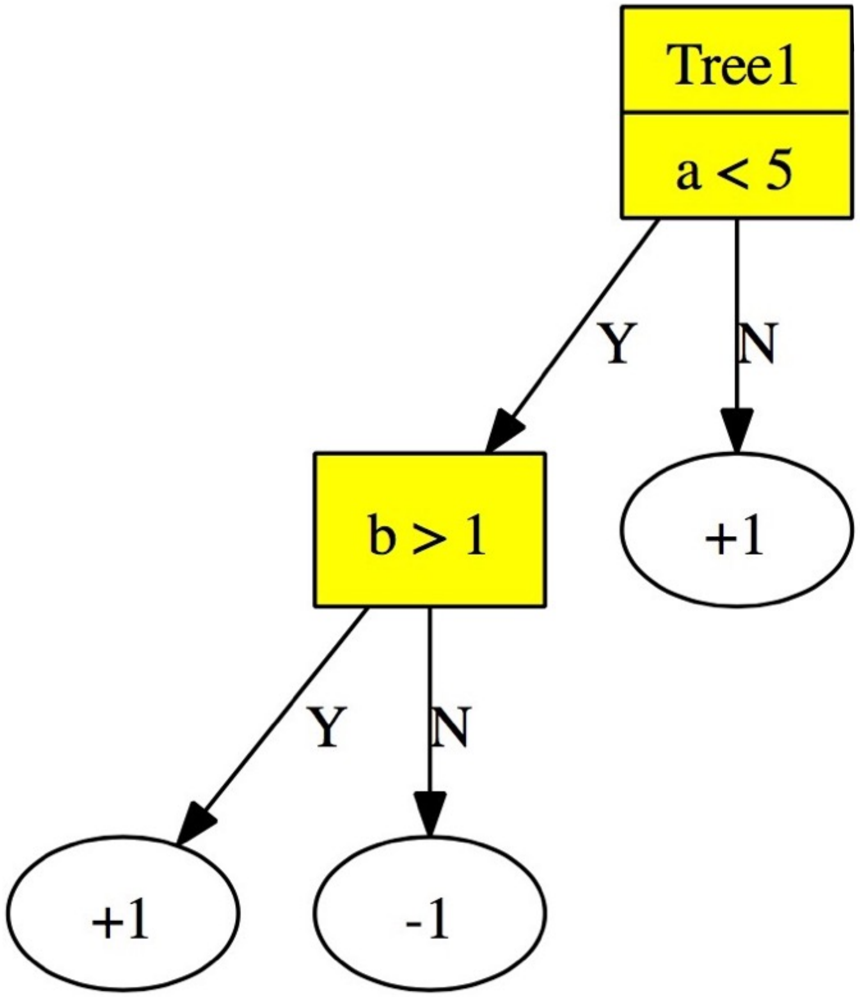
What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.
- Any predictor can be used as a base predictor.
  - In this talk, and in Spark, the base predictors are decision trees.

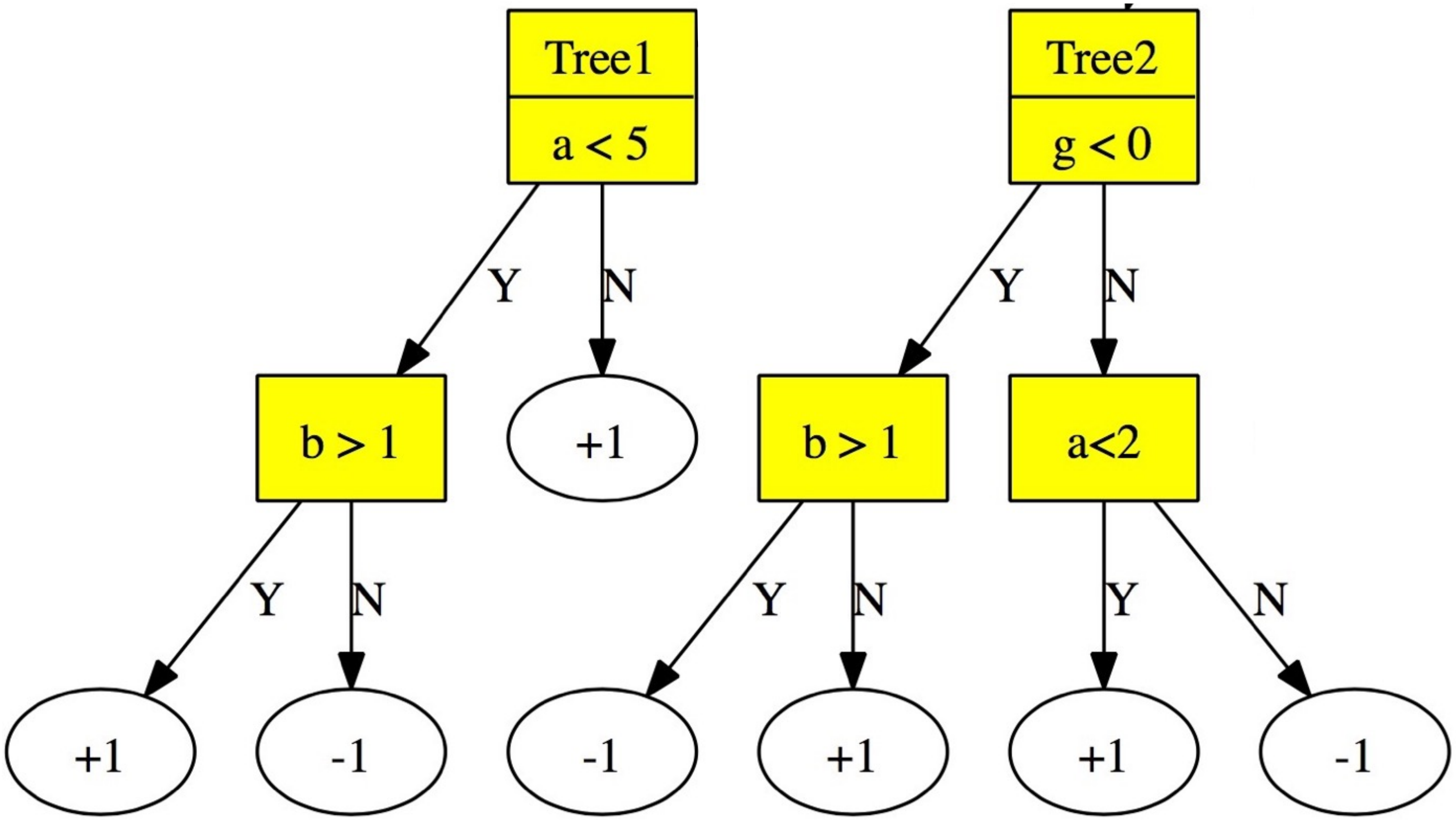
What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.
- Any predictor can be used as a base predictor.
  - In this talk, and in Spark, the base predictors are decision trees.
  - We will restrict our attention to binary classification, but there are solutions for multi class and for regression.

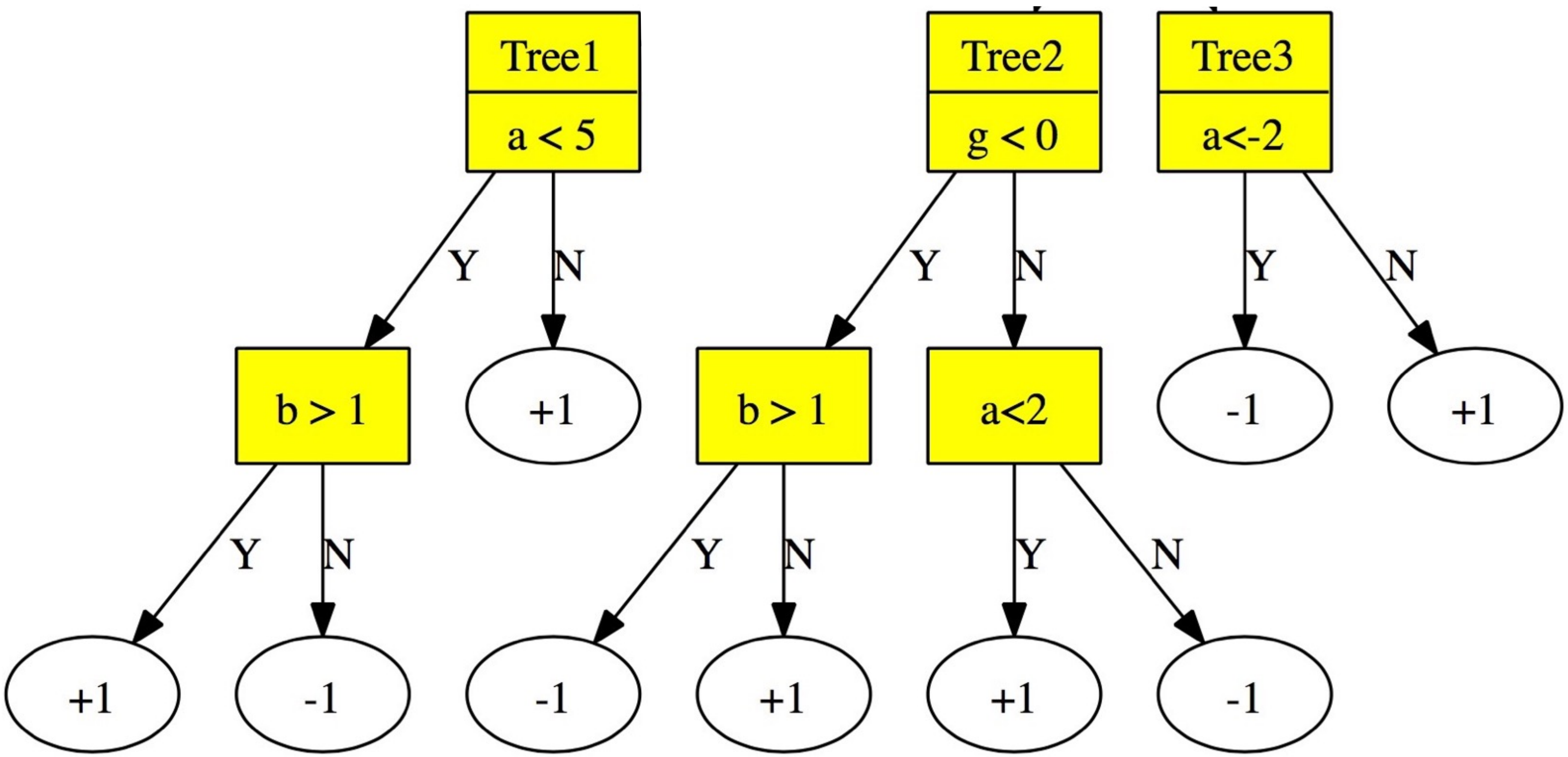
An Ensemble of trees



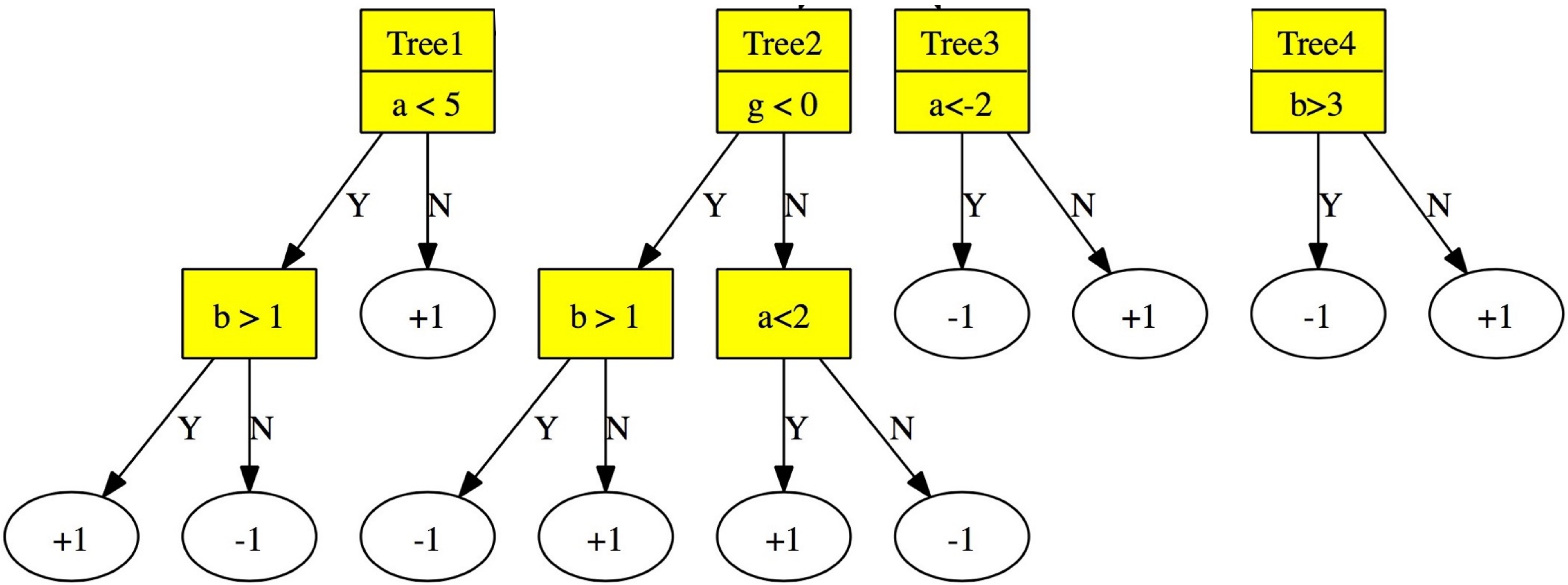
An Ensemble of trees



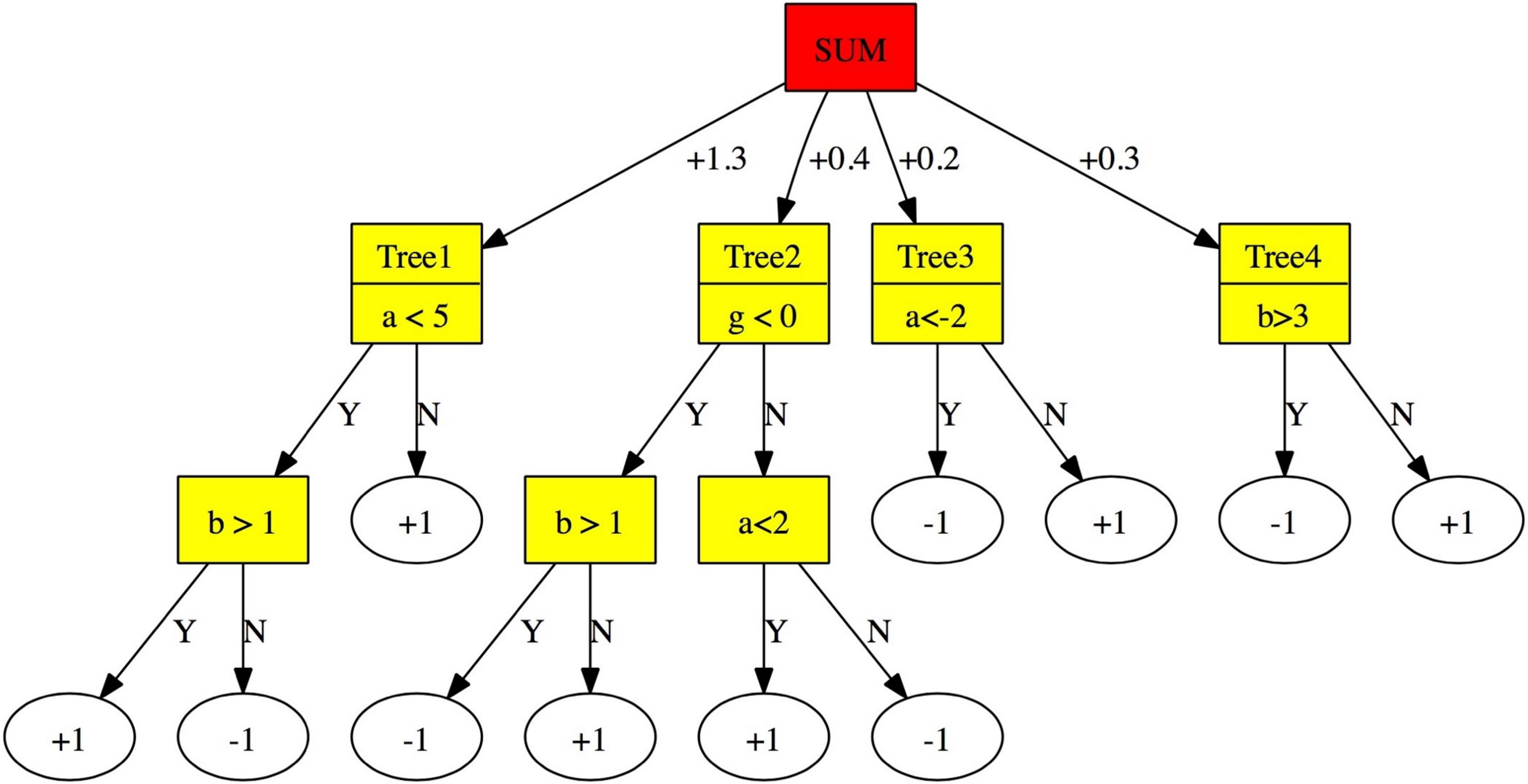
An Ensemble of trees



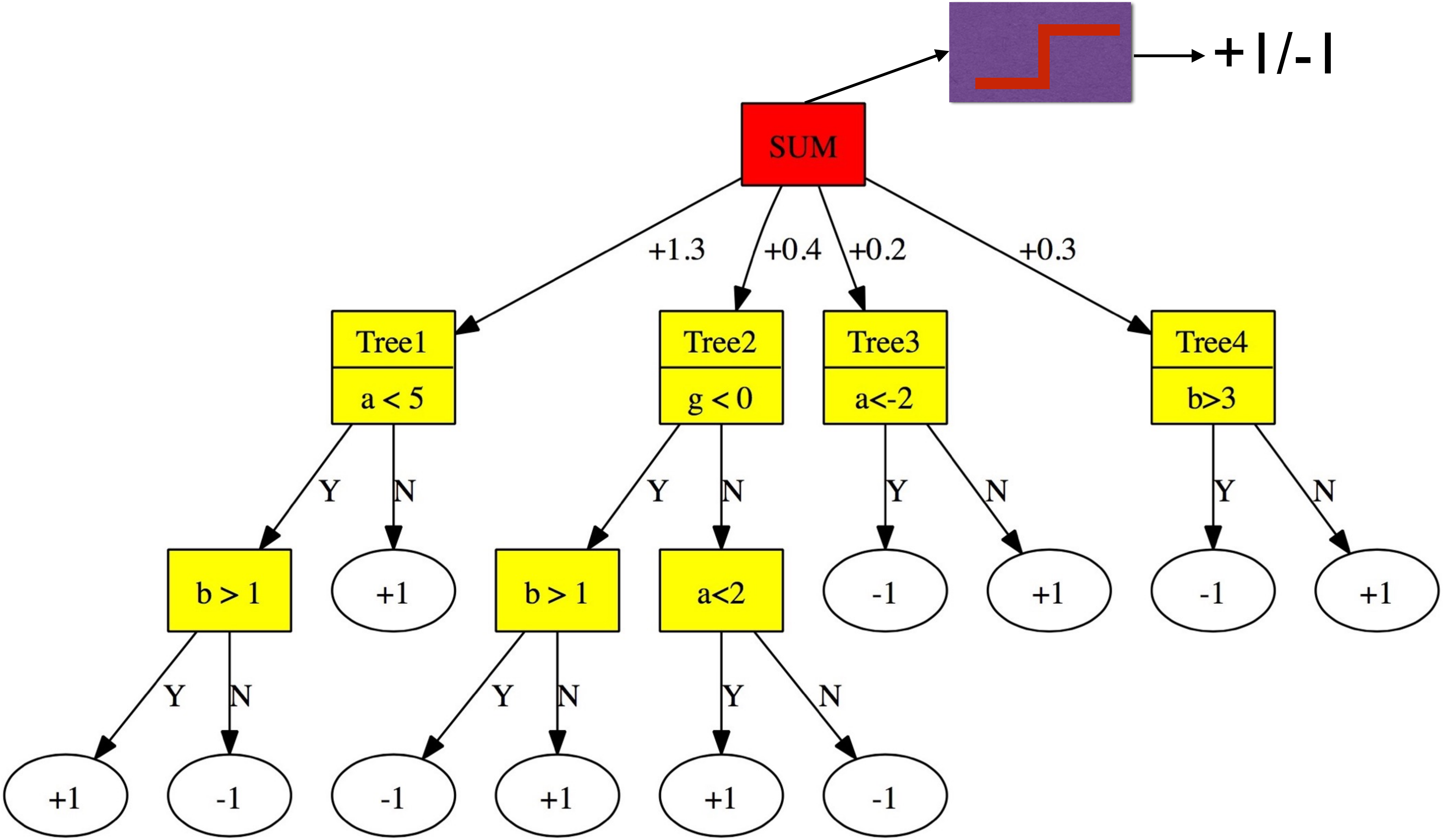
An Ensemble of trees



An Ensemble of trees



An Ensemble of trees



Bagging = bootstrap aggregation

Bagging = bootstrap aggregation

- Decision trees have high data variation.

Bagging = bootstrap aggregation

- Decision trees have high data variation.
    - i.e. the generated tree is sensitive to small changes in the training set.

Bagging = bootstrap aggregation

- Decision trees have high data variation.
  - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.

Bagging = bootstrap aggregation

- Decision trees have high data variation.
    - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.
    - Running an algorithm over random resampling is called “The Bootstrap”

Bagging = bootstrap aggregation

- Decision trees have high data variation.
  - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.
  - Running an algorithm over random resampling is called "The Bootstrap"
- Trees can be learned in parallel and

Bagging = bootstrap aggregation

- Decision trees have high data variation.
  - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.
  - Running an algorithm over random resampling is called “The Bootstrap”
- Trees can be learned in parallel and
- The result is a reduction in variation with no increase

Random Forests

Random Forests

- Based on bagging trees.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.
- Speeds up the learning process.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.
- Speeds up the learning process.
- All trees get equal weight (1.0)

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.
- Speeds up the learning process.
- All trees get equal weight (1.0)
- All trees can be learned in parallel.

Gradient Tree Boosting

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.
- Each tree is trained using a **weighted** training set. The weights represent the gradient of the loss function.

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.
- Each tree is trained using a **weighted** training set. The weights represent the gradient of the loss function.
- Each tree receives a different weight (corresponding to the alpha in adaboost)

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.
- Each tree is trained using a **weighted** training set. The weights represent the gradient of the loss function.
- Each tree receives a different weight (corresponding to the alpha in adaboost)
- Stochastic gradient boosting: use random resampling of the training set a.k.a. Bagging.