Using Ensembles of Algorithms to Overcome Overfitting

Overview

Understand the problem of overfitting and it's causes

Understand how to overcome overfitting in decision trees

Understand how to use Ensemble Learning to overcome overfitting

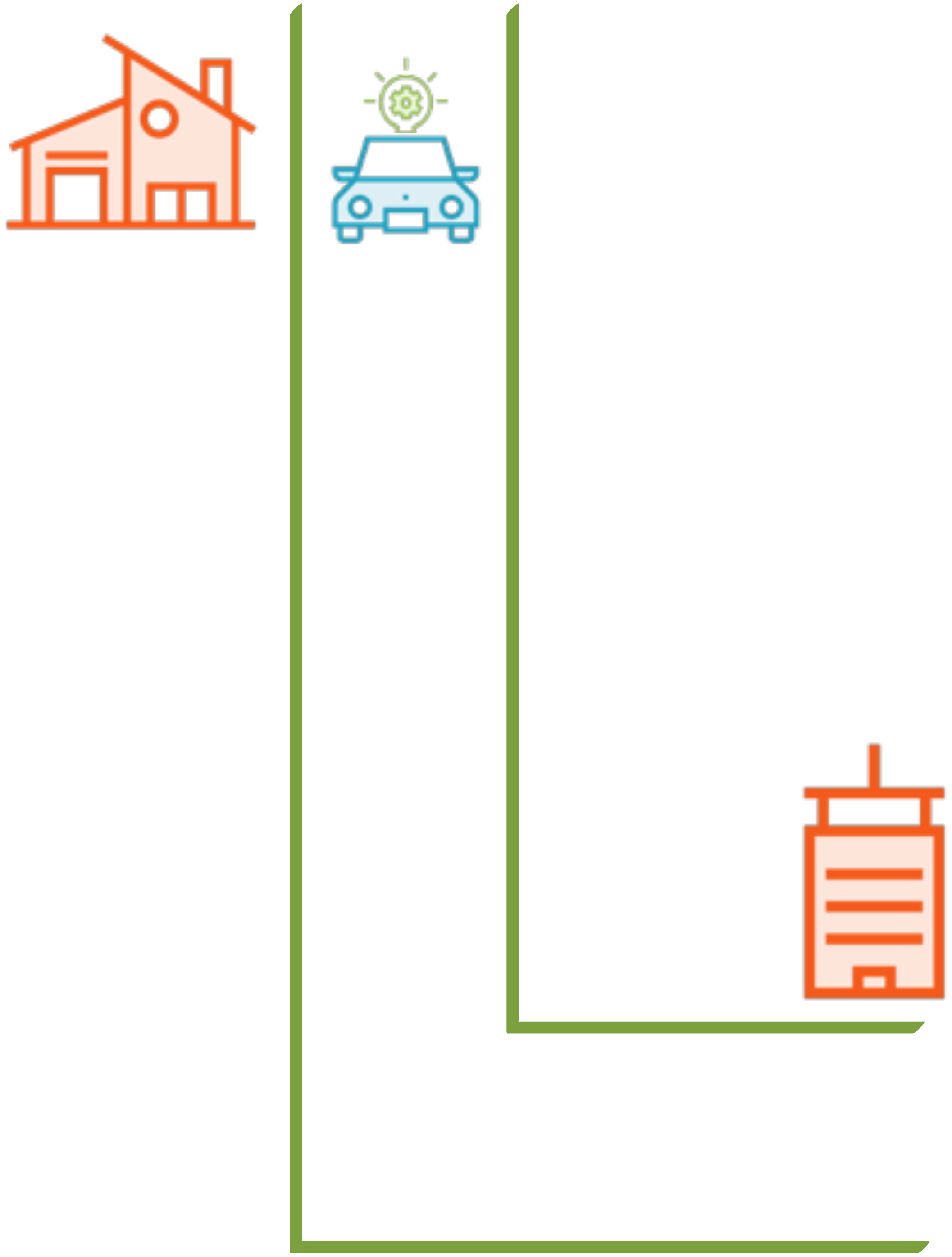
Machine Learning

A computer program/system

that can learn from

“Experience”

Teaching a Route to a Car



Drive 1 mile south

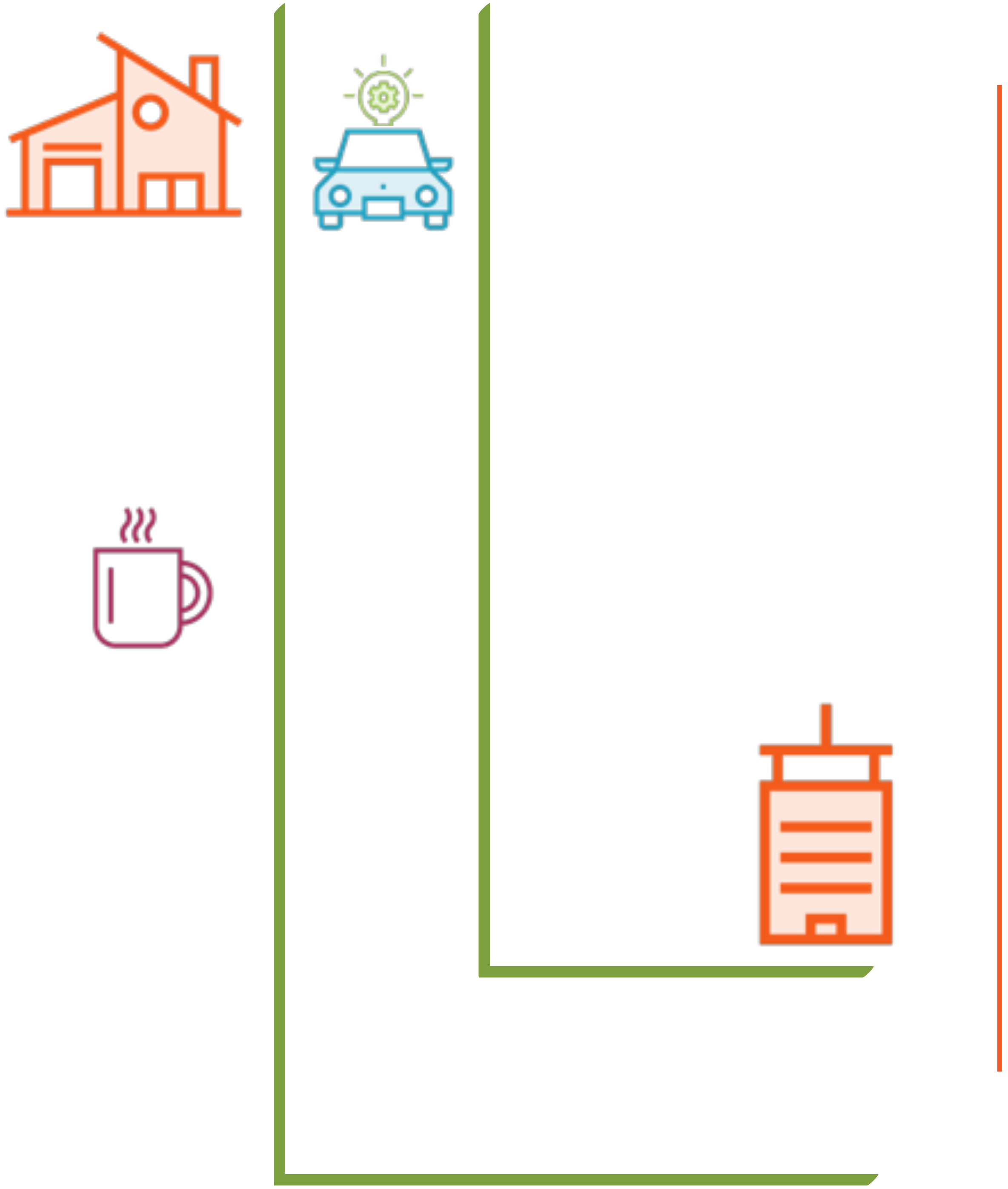
Stop

Turn left

Drive 0.5 mile east

Stop

Teaching a Route to a Car



Drive 0.5 mile south

Stop

Drive 0.5 mile south

Stop

Turn left

Drive 0.5 mile east

Stop

Machine Learning

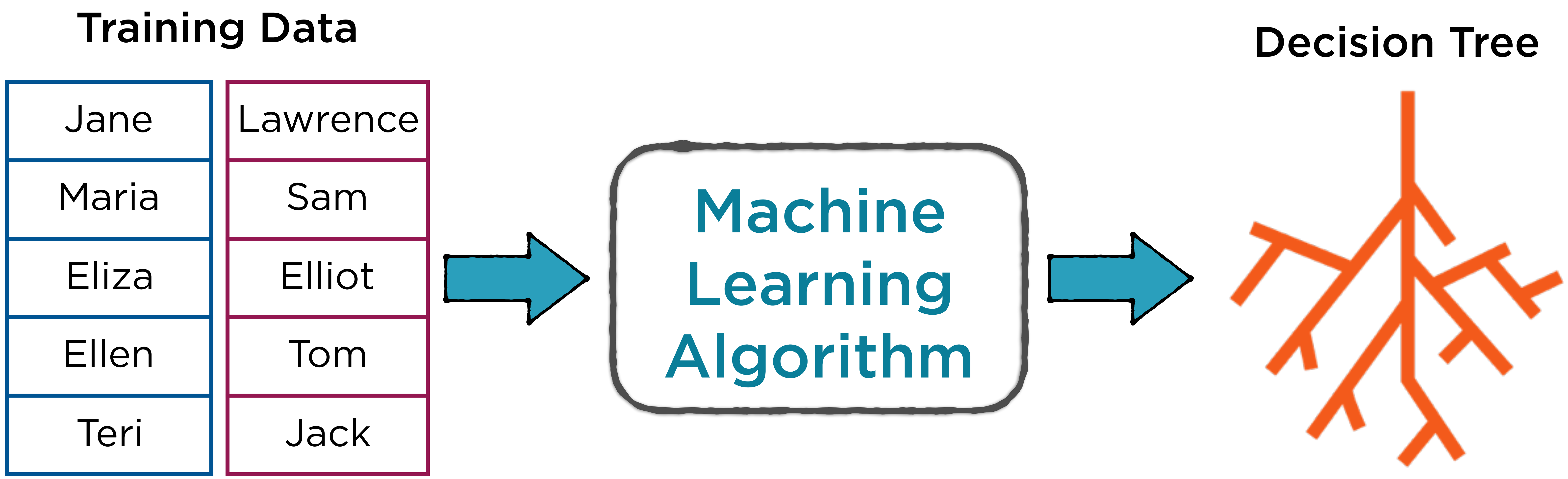
“Experience” ↔ Data

Gender Detection

Given the first name of a user

or

Build a Decision Tree



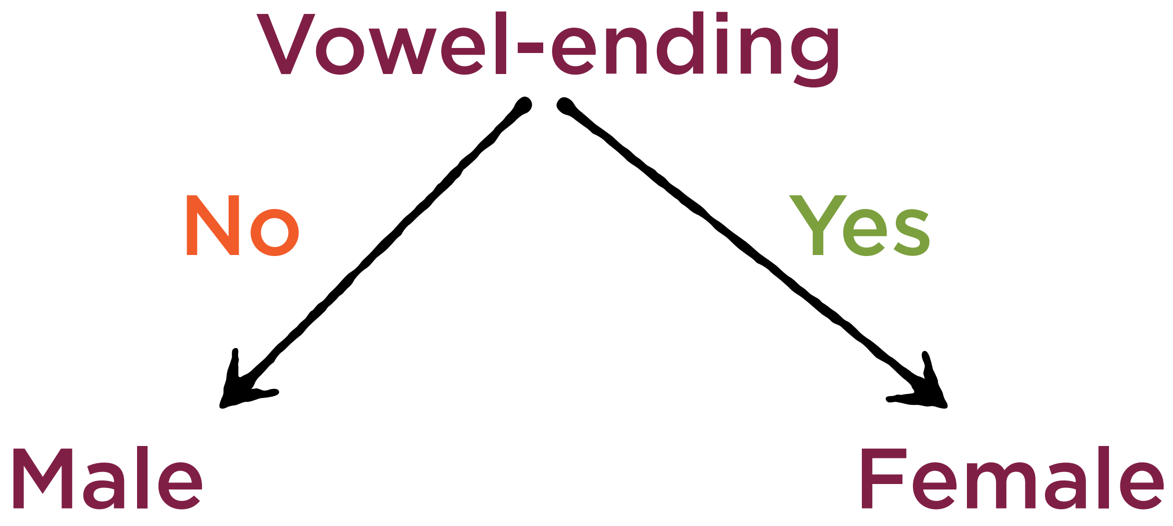
Training Data

Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

Machine Learning Algorithm

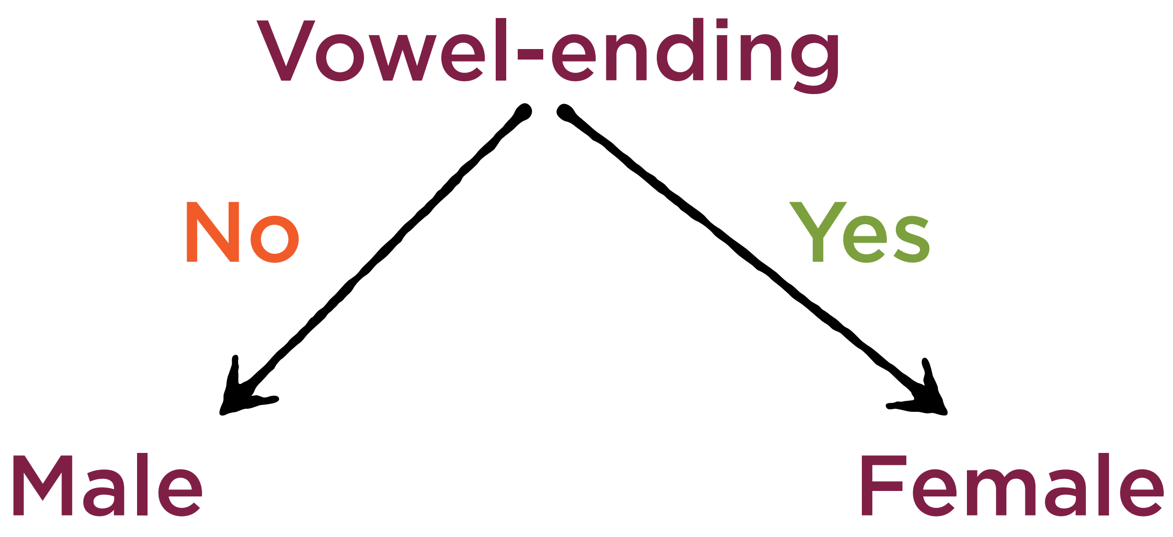
Decision Tree

Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

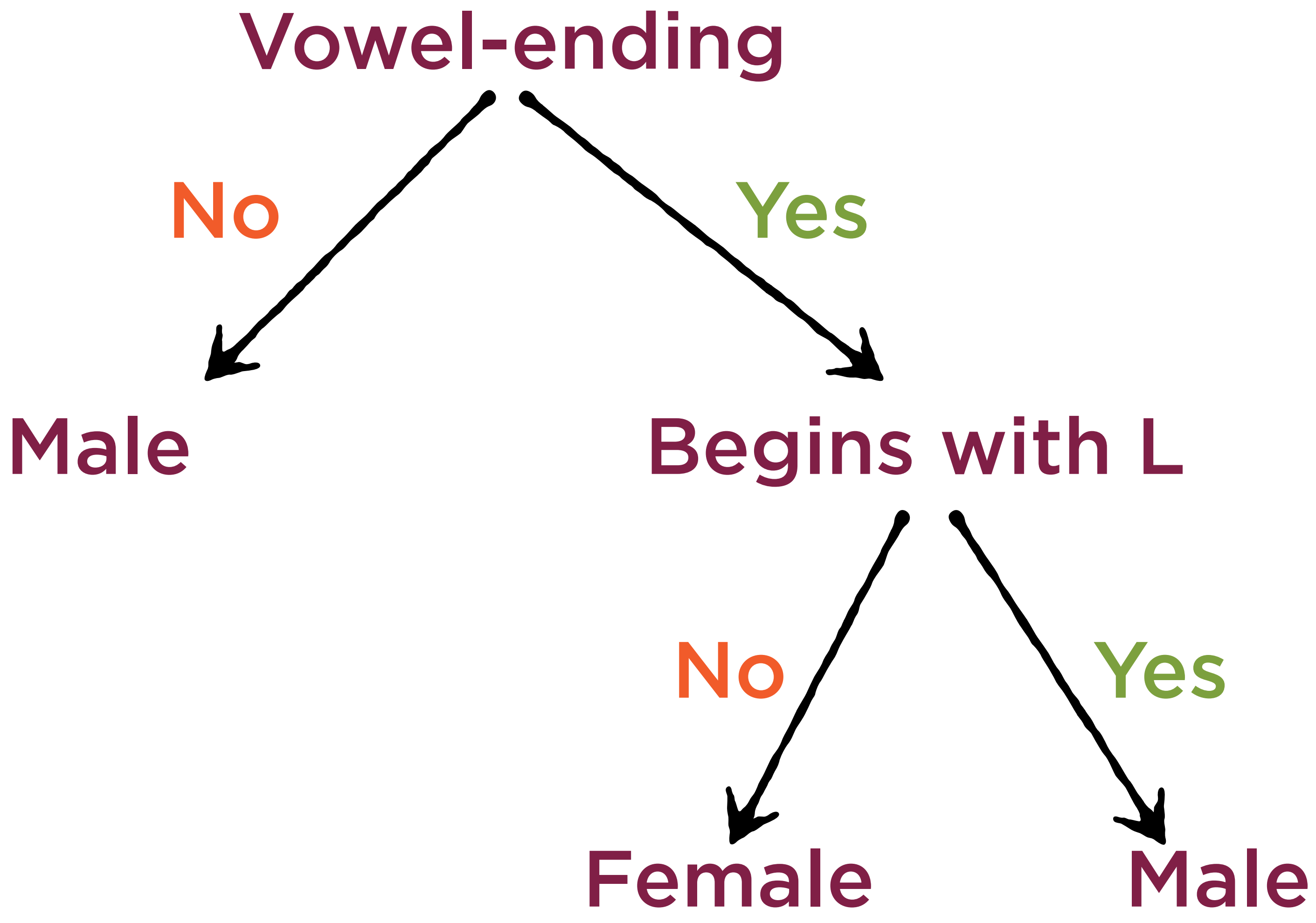


Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

80% accuracy on training data

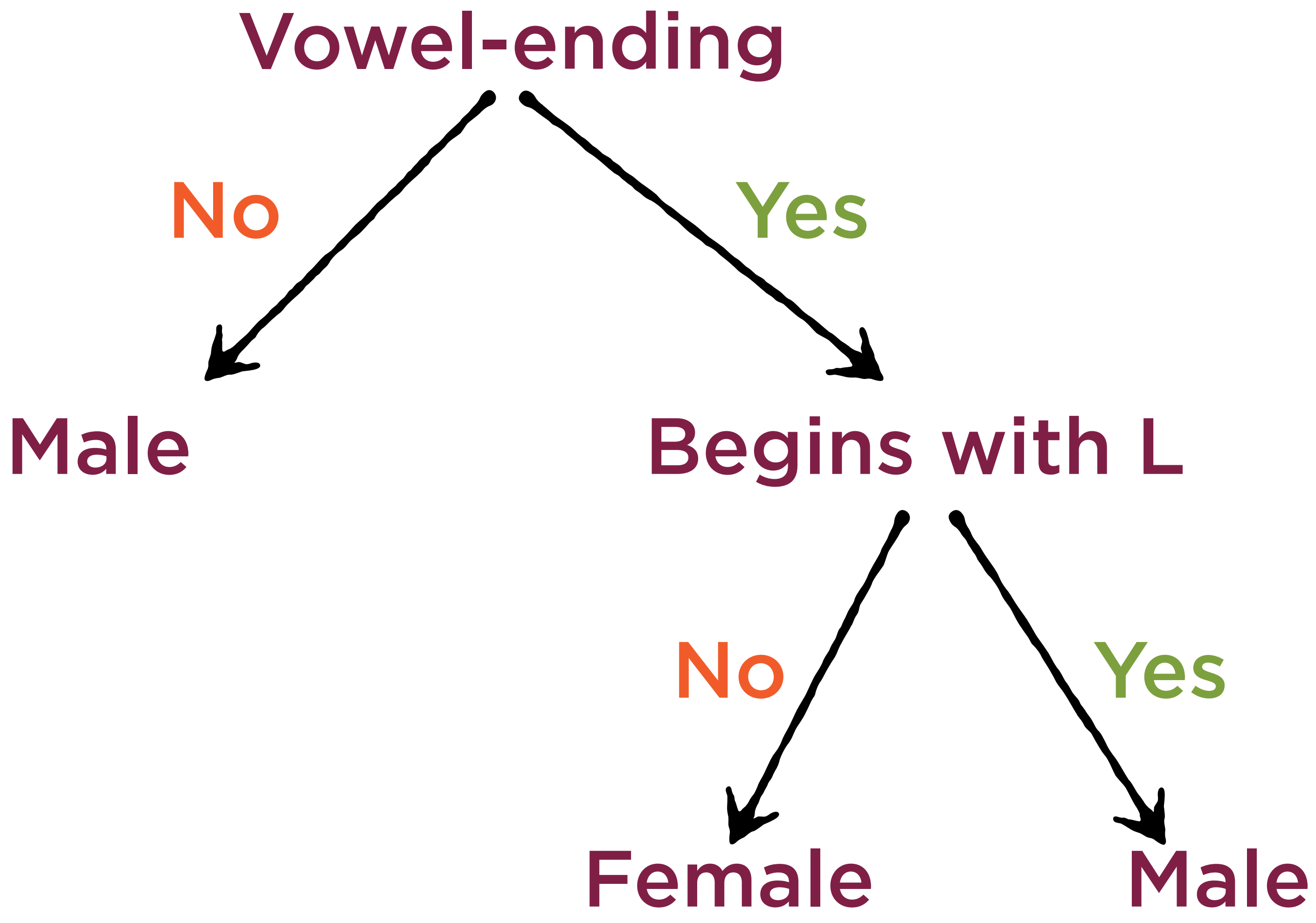


Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

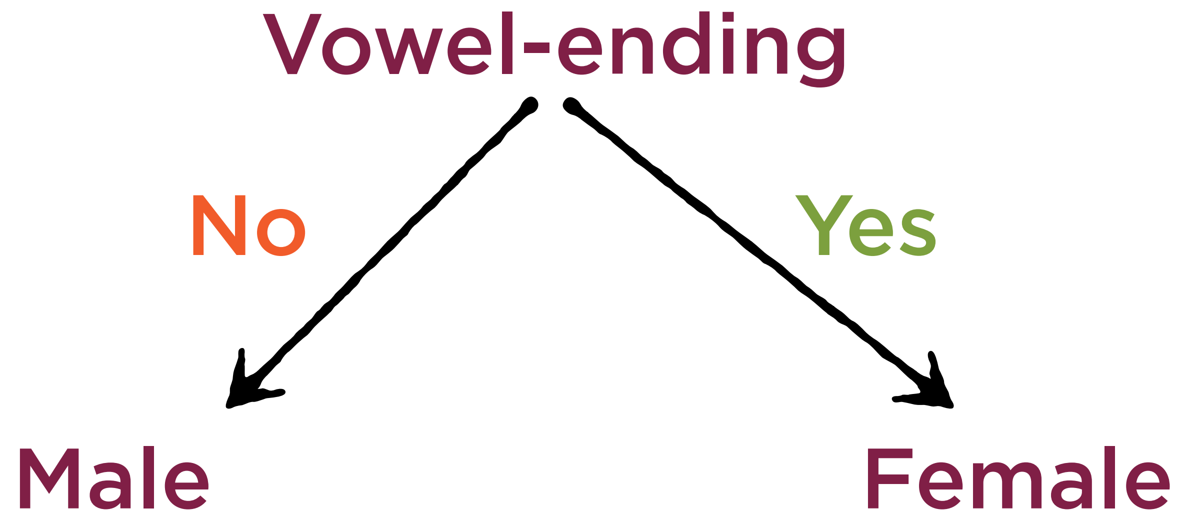


Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

90% accuracy on training data



Lyla	Mark
Radha	Robert
Jan	Jon



Vowel-ending

No

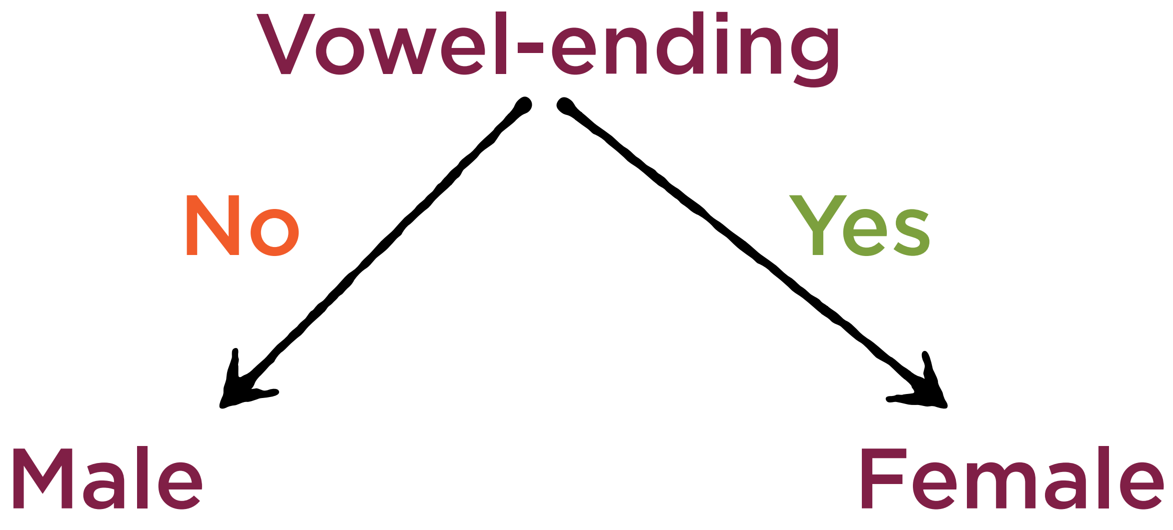
Yes

Male

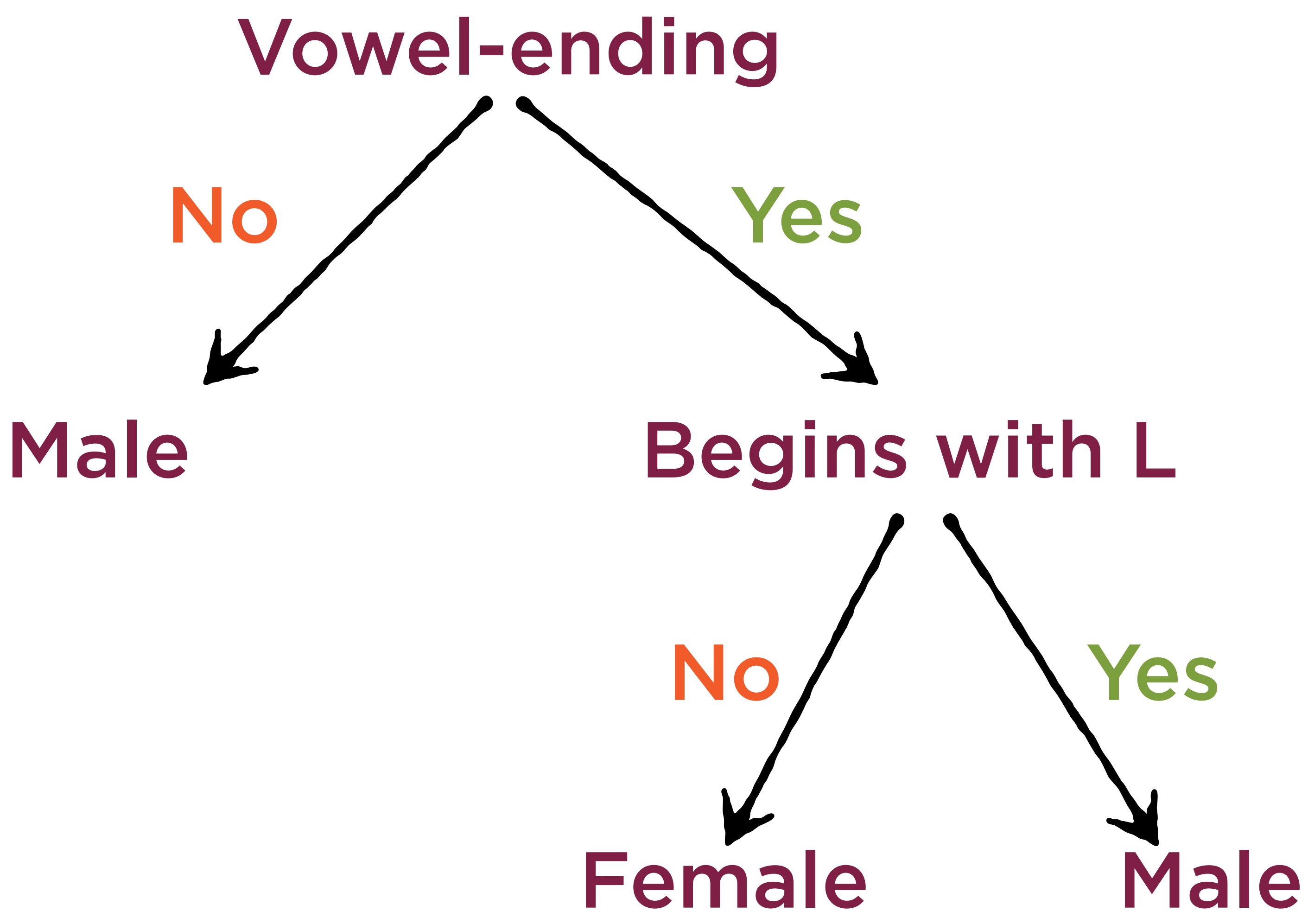
Female

Lyla	Mark
Radha	Robert
Jan	Jon

83% accuracy on test data



Lyla	Mark
Radha	Robert
Jan	Jon



Vowel-ending

No

Yes

Male

Begins with L

No

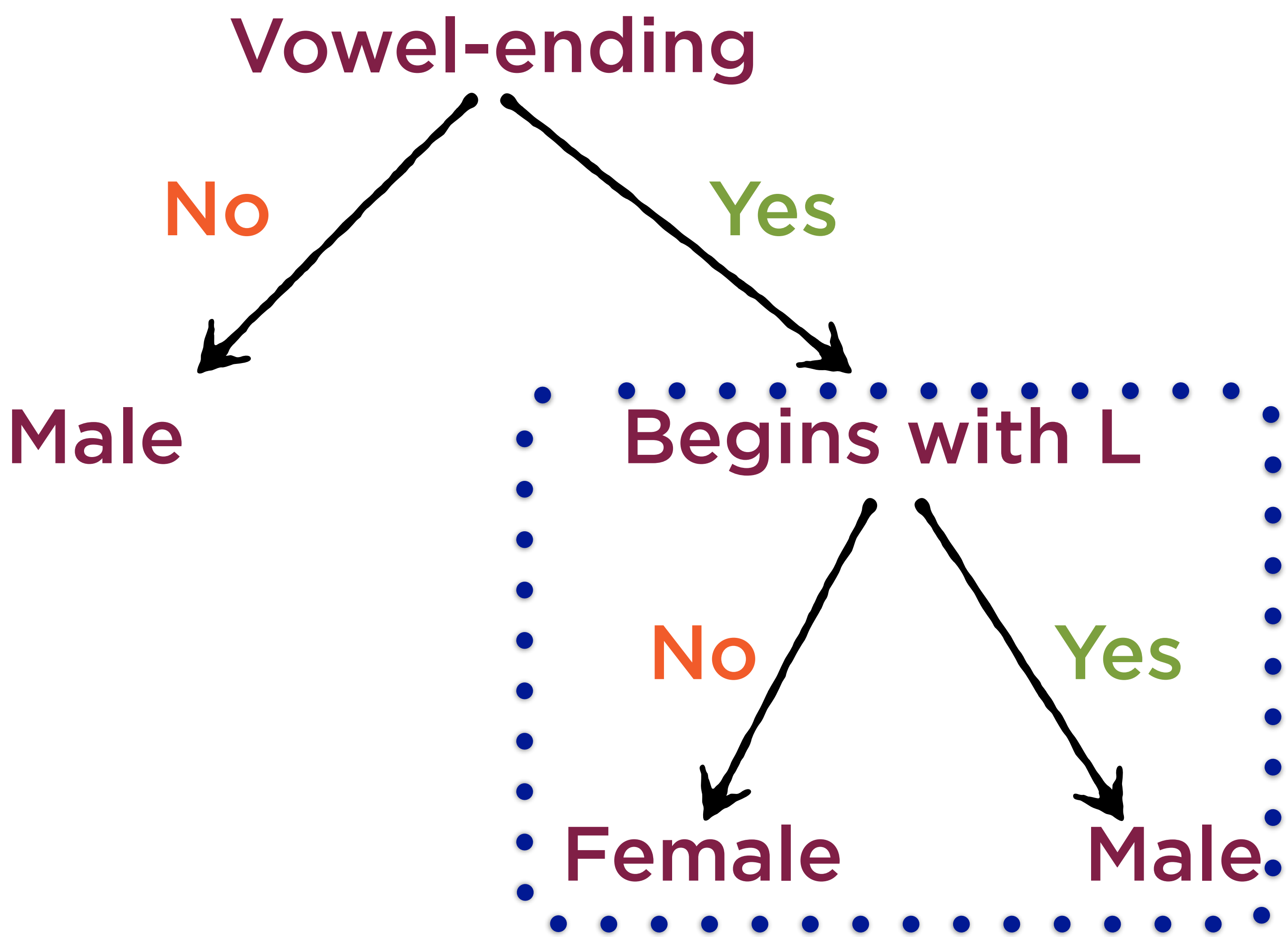
Yes

Female

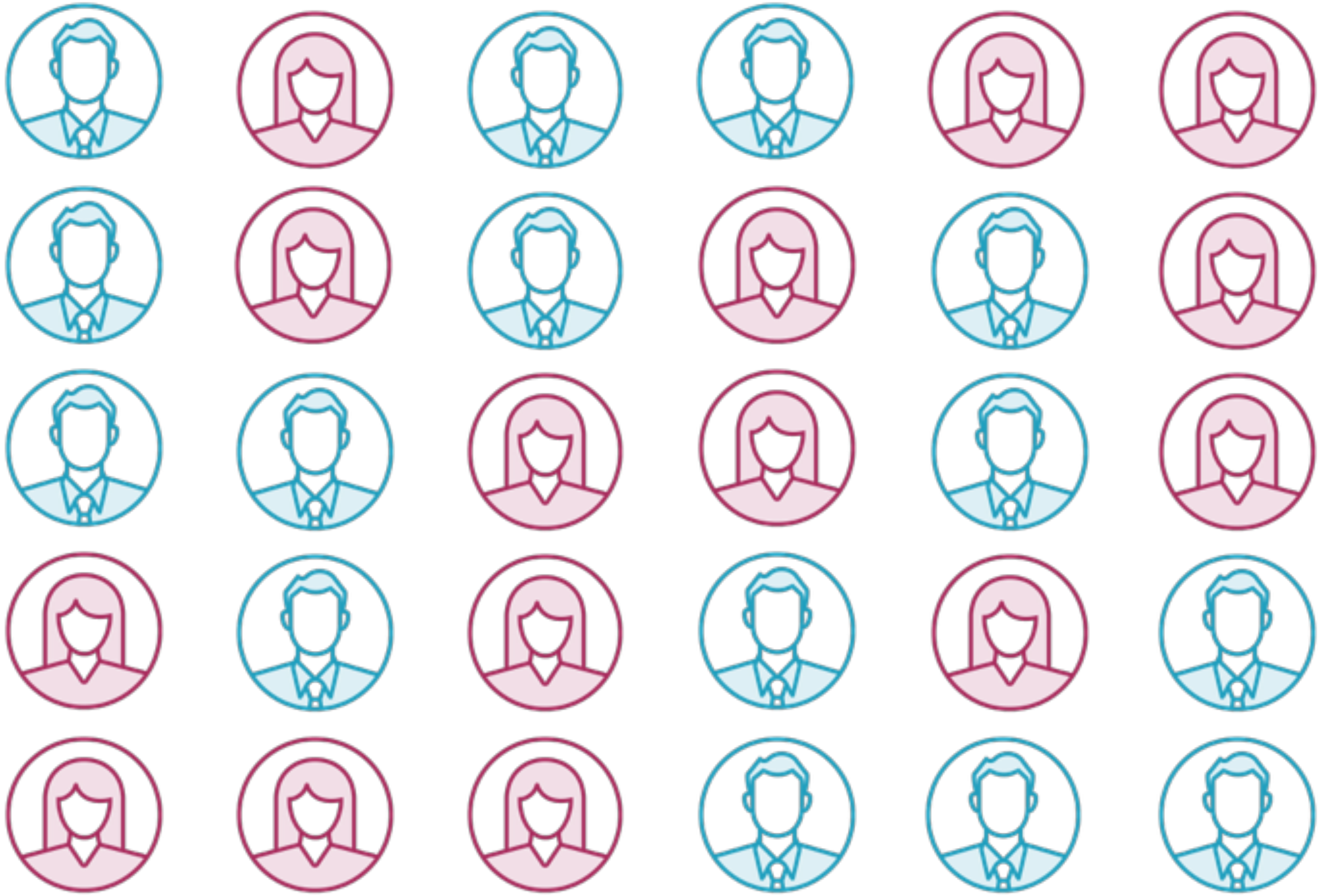
Male

Lyla	Mark
Radha	Robert
Jan	Jon

67% accuracy on test data

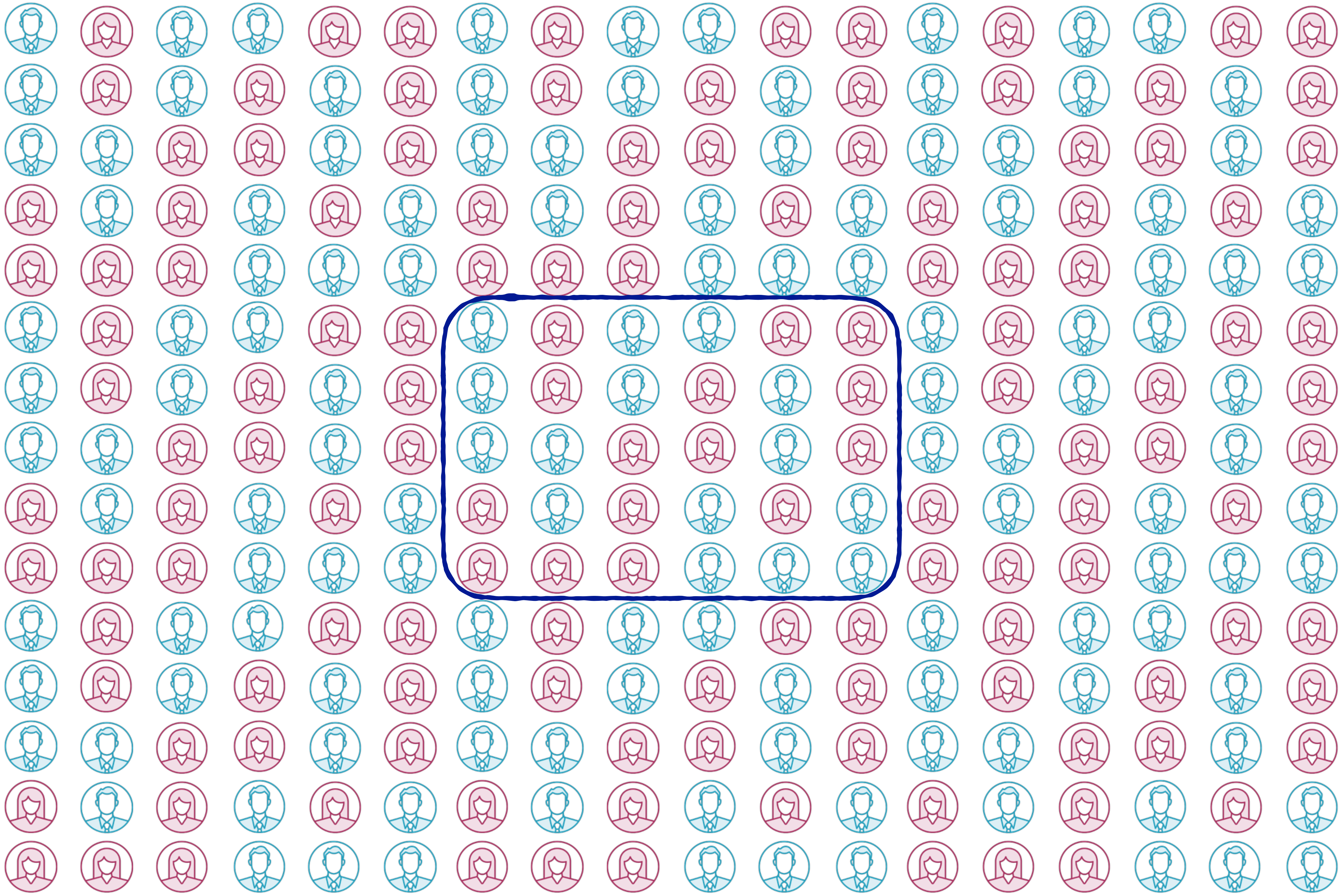


How Decision Tree Learning Works



Learn patterns from training data

How Decision Tree Learning Works



Learn patterns from training data

...that apply to the universe of data

Overfitting

Learning patterns that are irrelevant to the universe of data

Underfitting

Missing patterns that are relevant to the universe of data

Avoiding Overfitting

Pruning

Reduce complexity of a decision tree

Ensemble Learning

Build multiple decision trees and combine their results

Avoiding Overfitting

Pruning

Reduce complexity of a decision tree

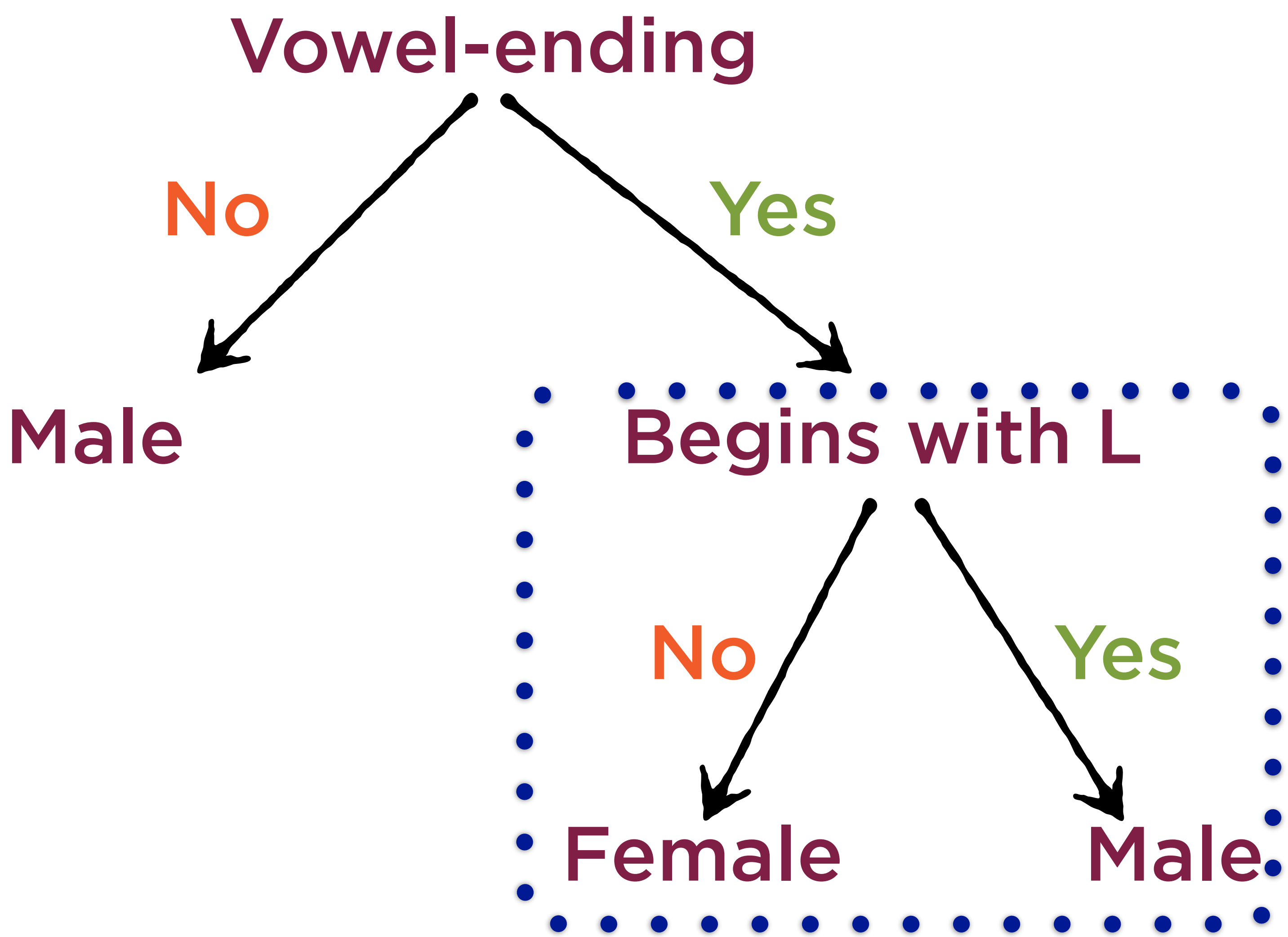
Ensemble Learning

Build multiple decision trees and combine their results

Pruning

Remove some of the nodes in your decision tree

Pruning



Vowel-ending

No

Yes

Male

Begins with L

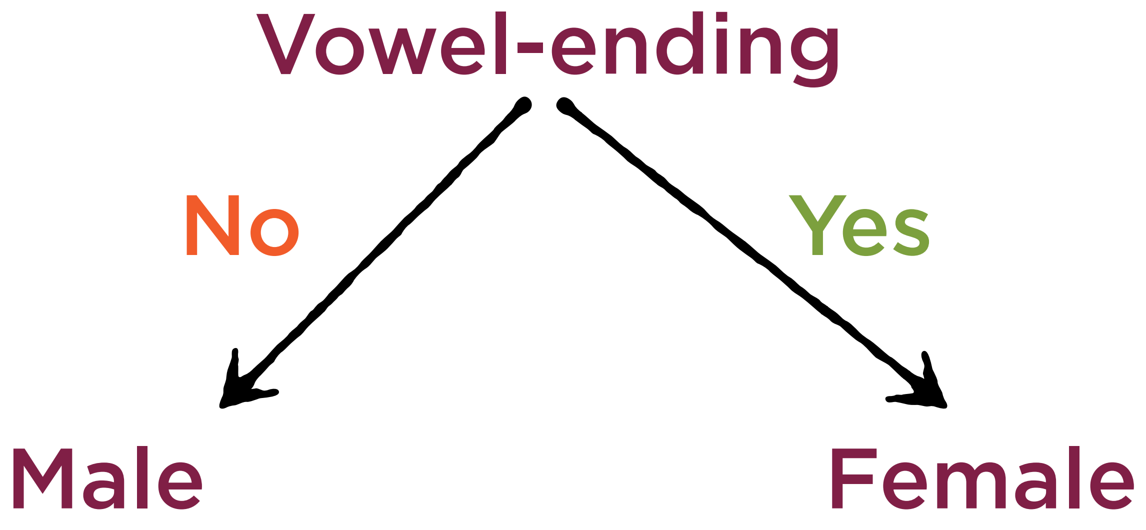
No

Yes

Female

Male

Pruning



Do this if accuracy on a test data set is not affected

Avoiding Overfitting

Pruning

Reduce complexity of a decision tree

Ensemble Learning

Build multiple decision trees and combine their results

Avoiding Overfitting

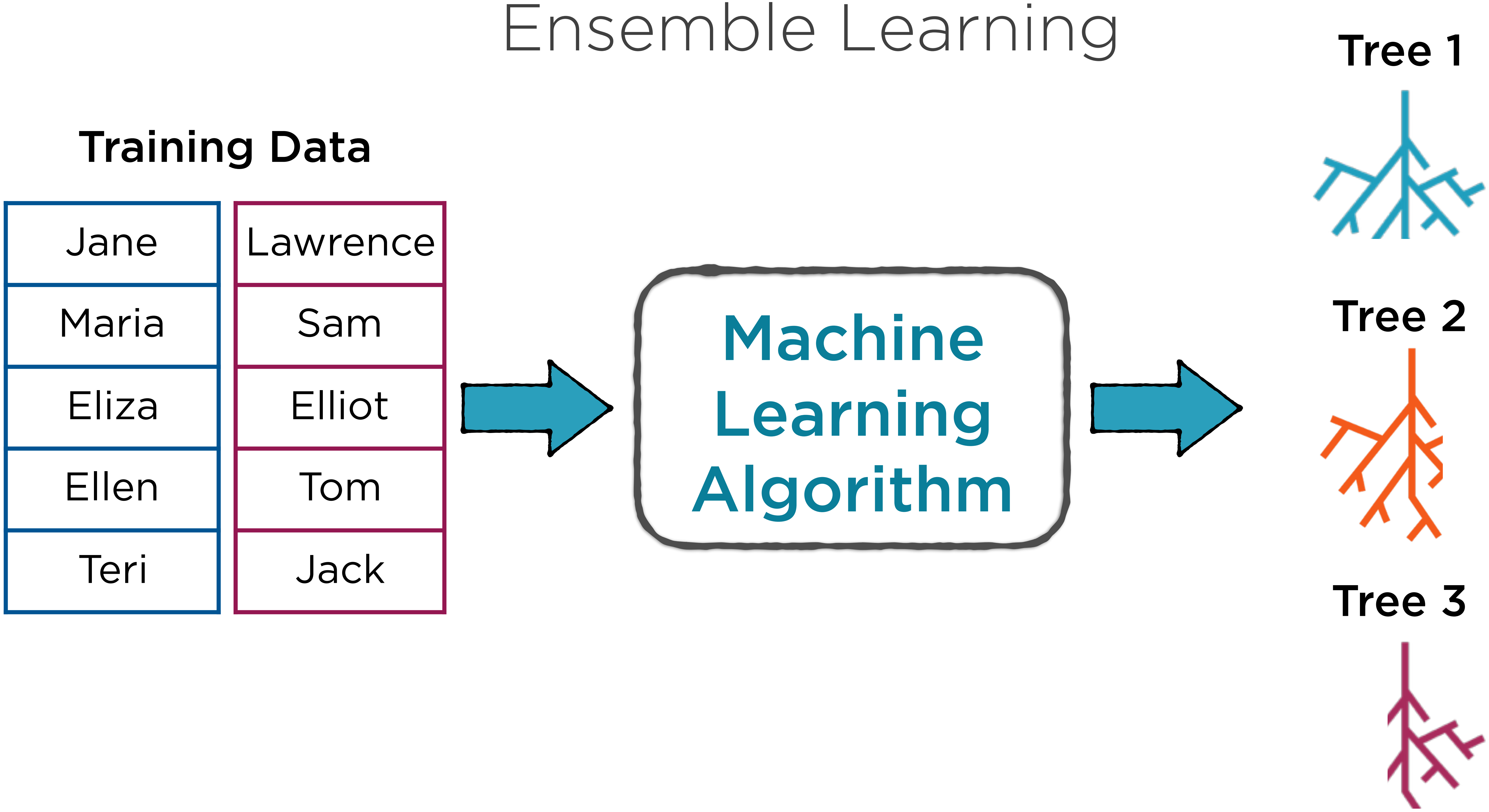
Pruning

Reduce complexity of a decision tree

Ensemble Learning

Build multiple decision trees and combine their results

Ensemble Learning



Training Data

Jane	Lawrence
Maria	Sam
Eliza	Elliot
Ellen	Tom
Teri	Jack

Machine Learning Algorithm

Ensemble Learning

Tree 1

Tree 2

Tree 3

Each tree will overfit to a different extent

Ensemble Learning

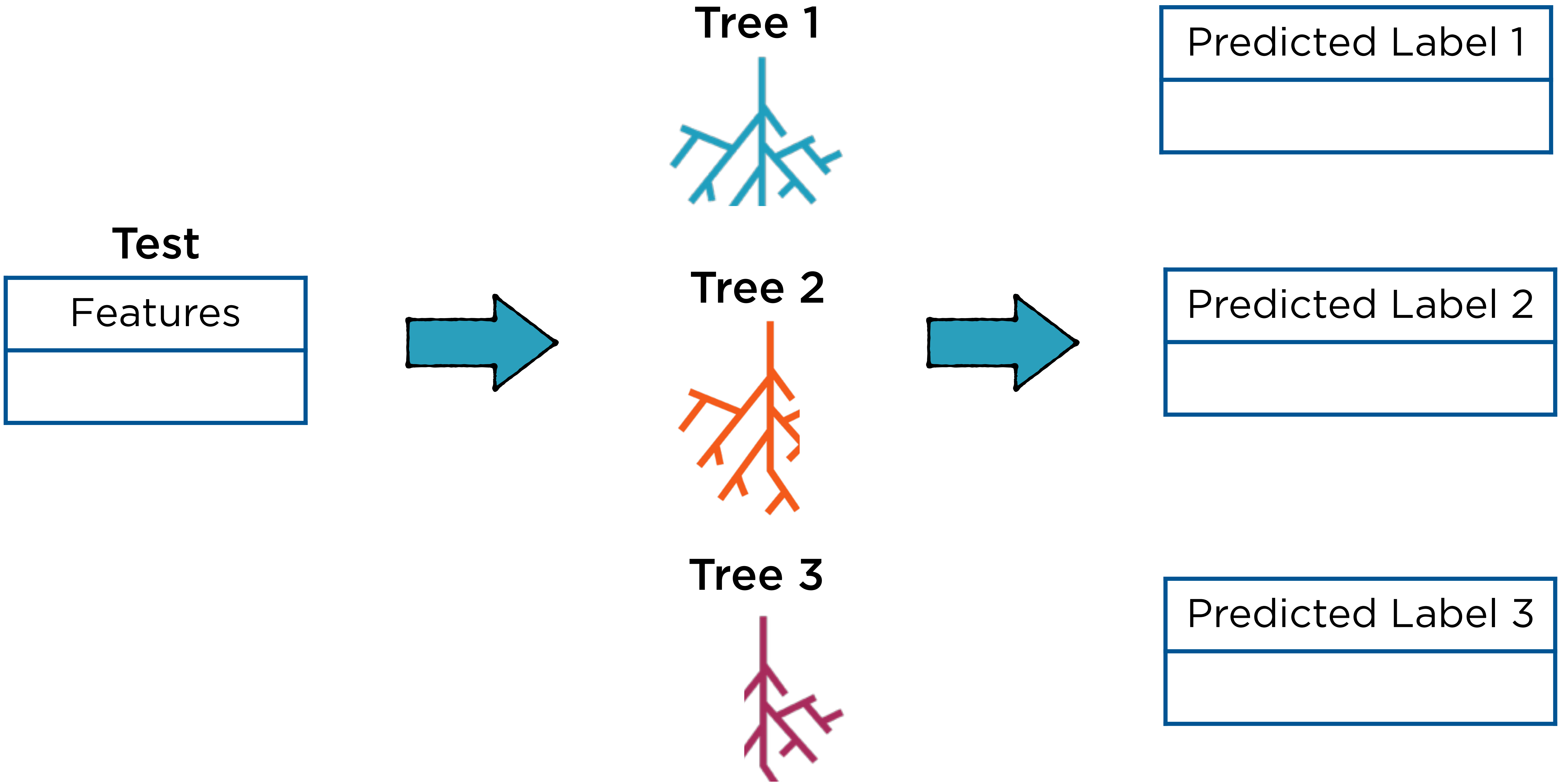
Tree 1

Tree 2

Tree 3

When you combine the results, the overfitting components cancel out

Ensemble Learning Test Phase



Test

Features

Tree 1

Tree 2

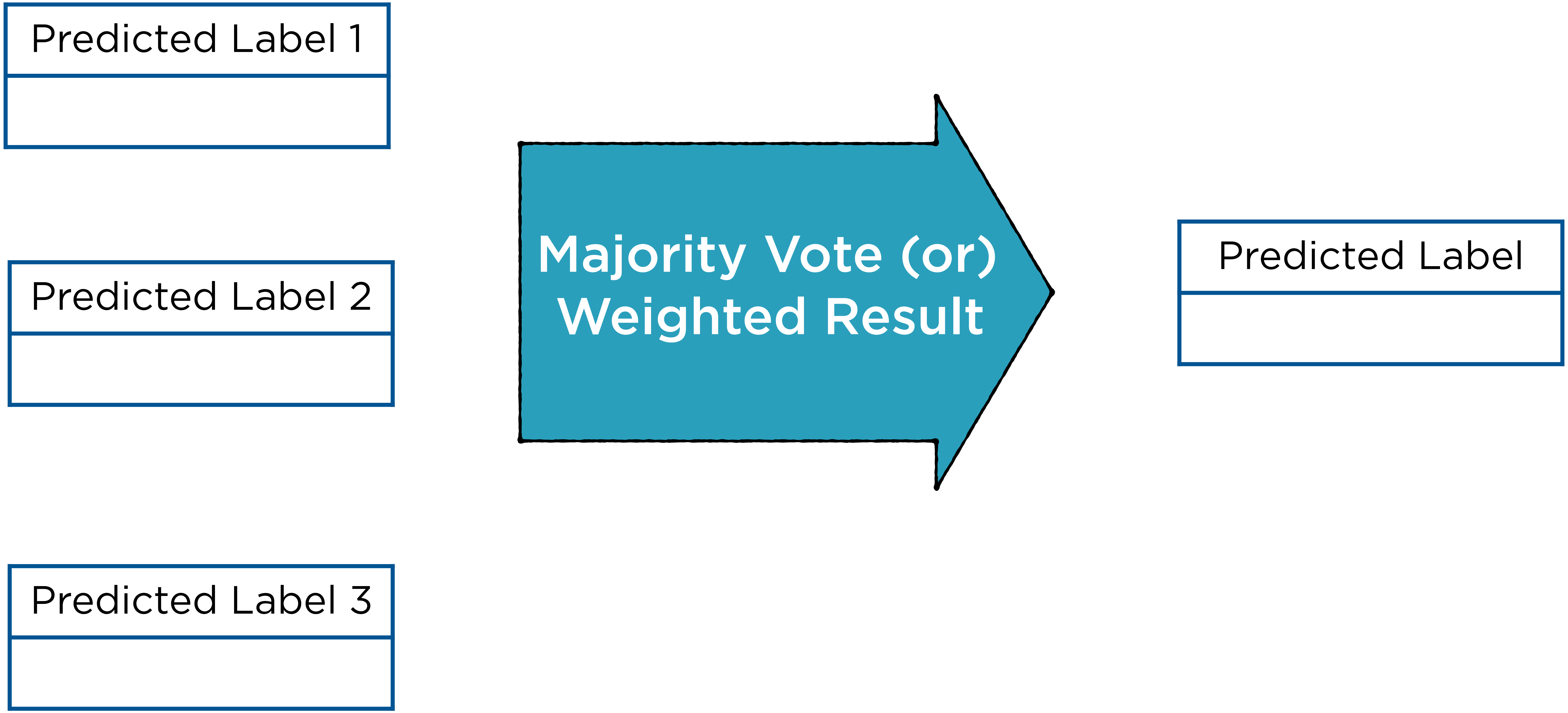
Tree 3

Predicted Label 1

Predicted Label 2

Predicted Label 3

Ensemble Learning Test Phase



Predicted Label 1

Predicted Label 2

Predicted Label 3

Majority Vote (or) Weighted Result

Predicted Label

Ensemble Learning

An ensemble is a collection of models

Ensemble Learning

Models built using different

Techniques

- Gini impurity vs Information gain
- Decision tree vs Logistic regression

Training Sets

Each tree built from a different subset of the training set

Features

Each tree built using a different subset of features

Parameters

Each tree built using different values of max tree depth

Ensemble Learning

Two techniques that use a combination of these 3

Training Sets

Each tree built from a different subset of the training set

Features

Each tree built using a different subset of features

Parameters

Each tree built using different values of max tree depth

Techniques

- Gini impurity vs Information gain
- Decision tree vs Logistic regression

Ensemble Learning

Random Forests

Each tree in the ensemble is built independently

Gradient Boosted Trees

Each tree is built with learnings from the previous tree

Summary

Understand the problem of overfitting and it's causes

Understand how to overcome overfitting in decision trees

Understand how to use Ensemble Learning to overcome overfitting