GENERAL ASSEMBLY

DECISION TREES AND RANDOM FORESTS

TODAY'S LEARNING OBJECTIVES

- Understand and build decision tree models for classification and regression with the sklearn library
- Understand and build random forest models for classification and regression
- Know how to extract the most important predictors in a random forest model

OPENING

DECISION TREES AND RANDOM FORESTS

I LOVE (CLASSIFYING) THE 90s

[Verse 1]

All right stop, collaborate and listen

Ice is back I got a brand new invention

Something grabs a hold of me tightly

Flow like a harpoon daily and nightly

Will it ever stop? Yo - I don't know

Now turn off the lights (huh) and I'll glow

And to the extreme I rock a mic like a vandal

Light up a stage and wax a chump like a candle

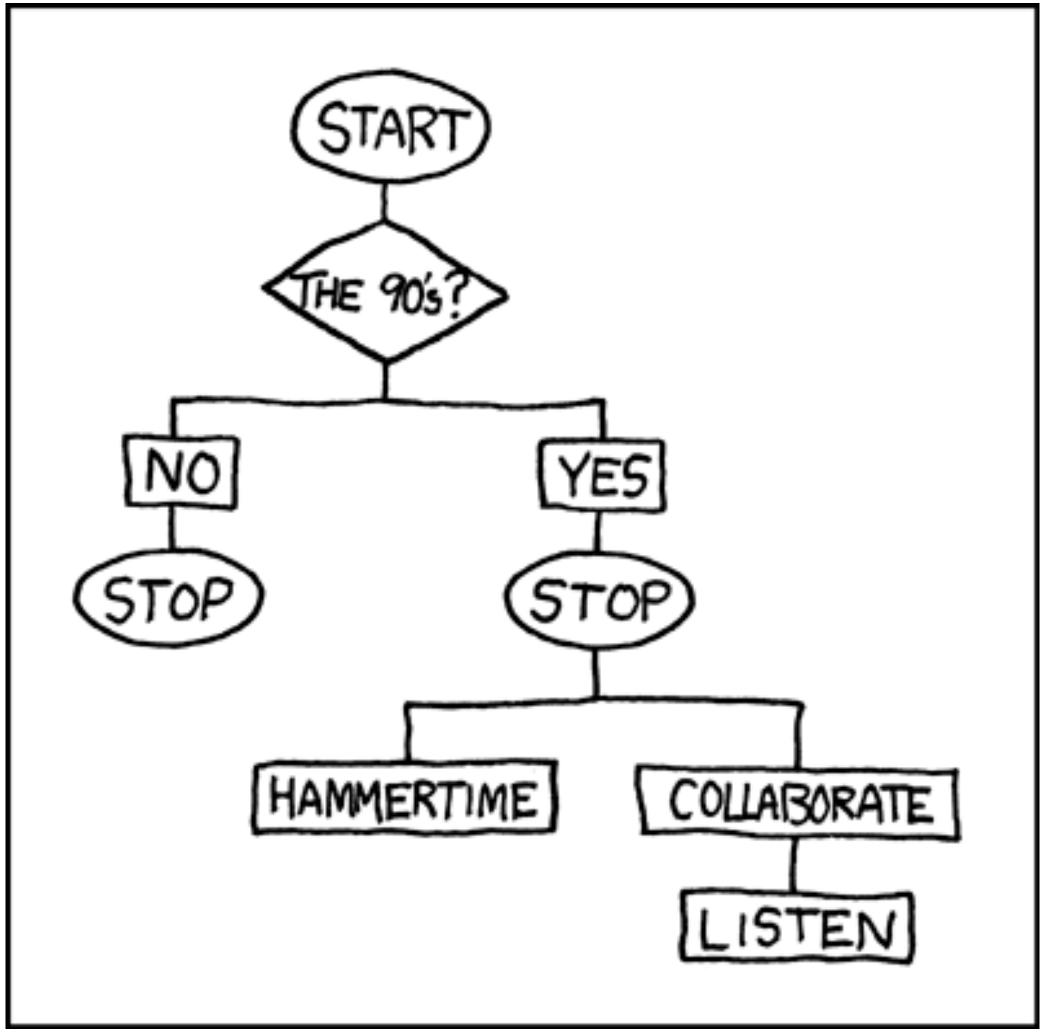
Too Cold – Vanilla Ice (1998)

[Breakdown]

Stop!

Hammer time

U Can't Touch This – MC Hammer (1990)

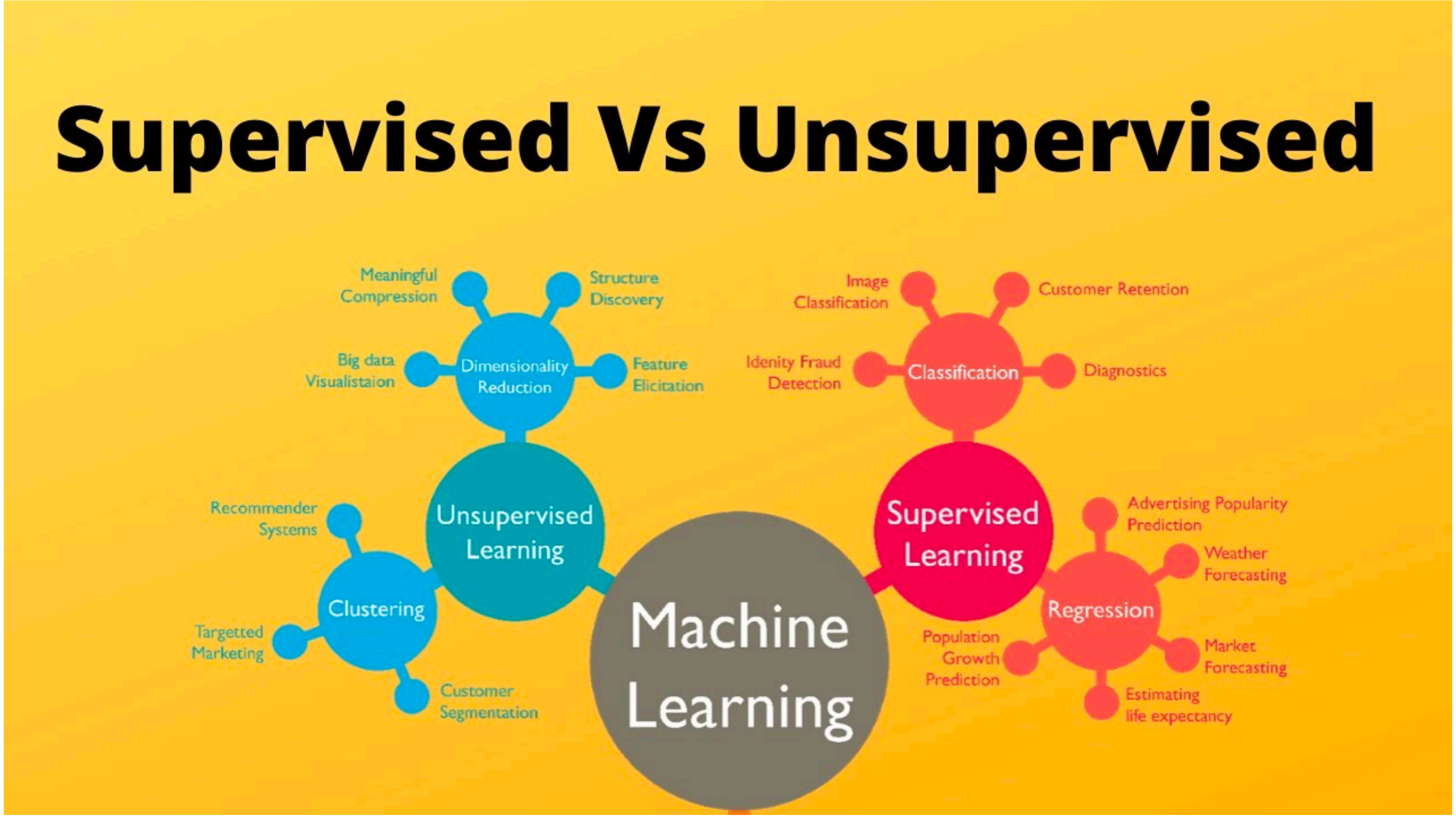


WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

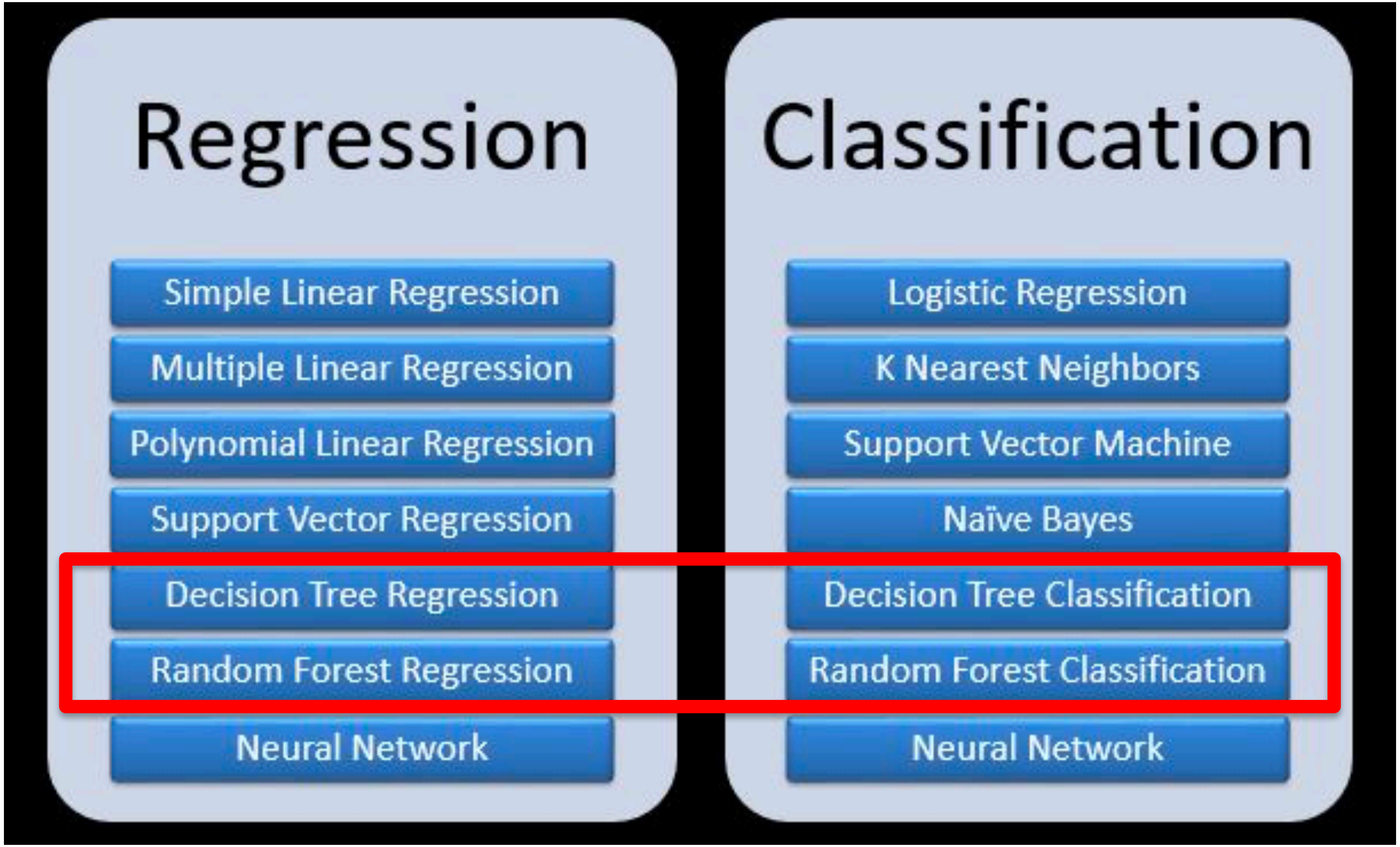
- Data has been **acquired** and **parsed.**

- Today we'll **refine** the data and **build** models (We'll also use plots to **represent** the results).

WHERE ARE WE IN THE MACHINE LEARNING UNIVERSE?



COMMON TREE ALGORITHMS



INTRODUCTION

DECISION TREES

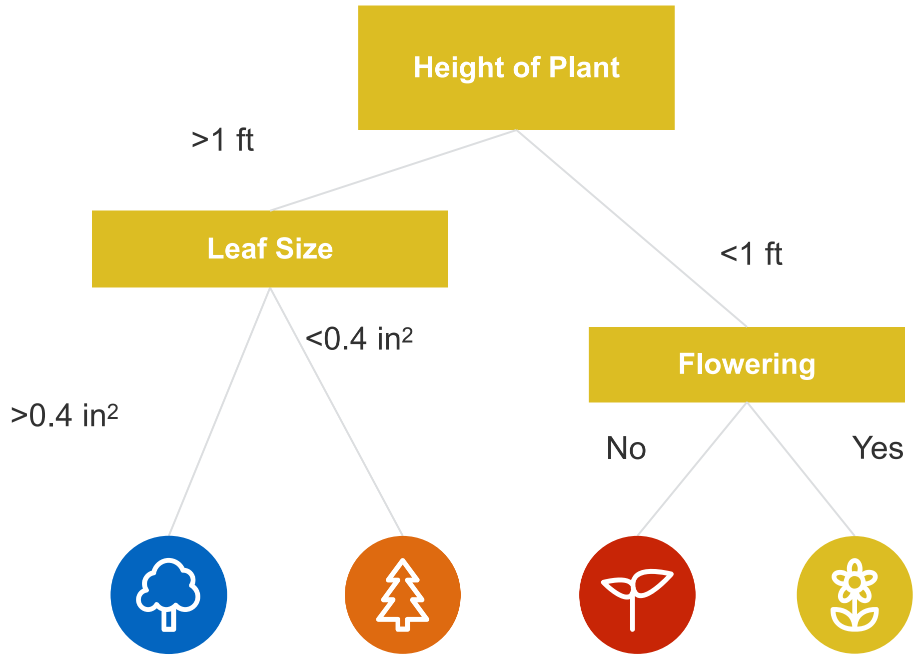
DECISION TREES

- Decision Trees are a machine learning model for regression and classification that develops *a series of yes/no rules* to explain the differences present in the outcome variable.

DECISION TREES

Decision trees use a machine learning algorithm which runs best in a supervised environment to recursively segment the data into subgroups that are as similar as possible with respect to the target.

Inputs	• Continuous and categorical variables
Outputs	• Regression • Classification
Strength	• Easily interpretable • Handle nonlinear relationships
Weakness	• Easy to over-fit without pruning • Sensitive to data changes

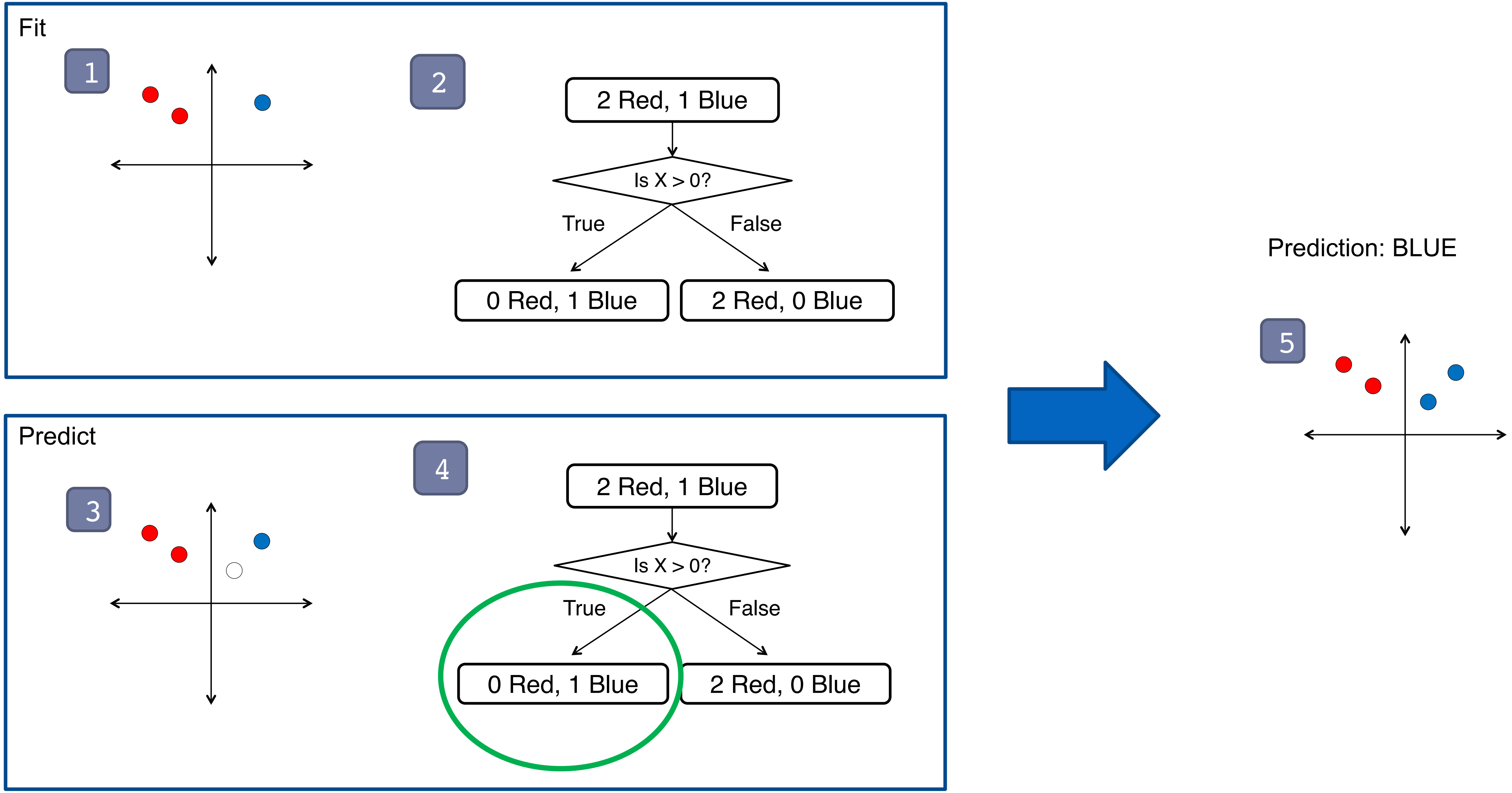


Based on height & leaf size – what kind of plant is this?

Do I have enough information to determine the type of plant?

Why is plant height more important than leaf size?

DECISION TREES

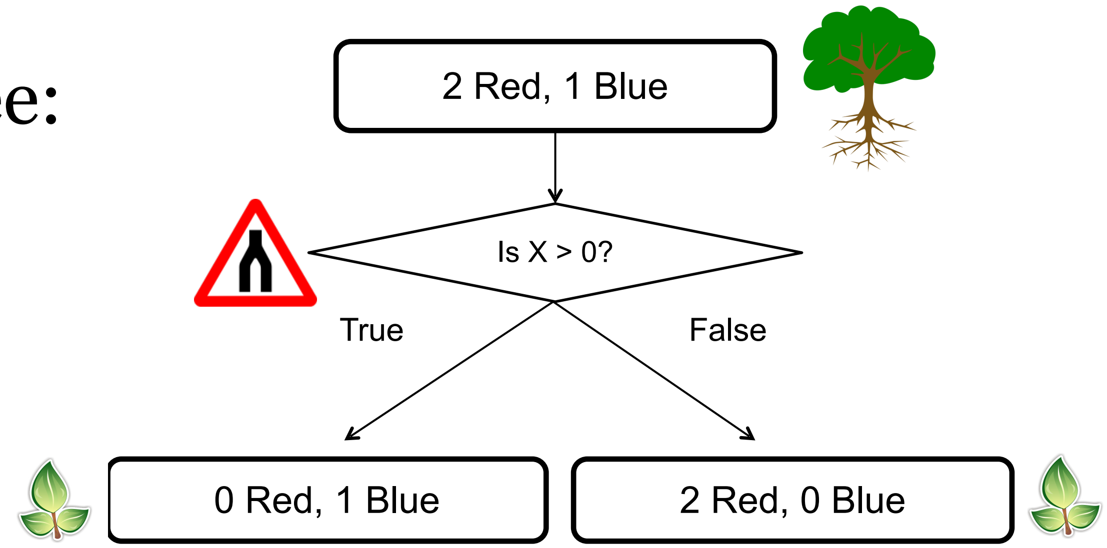


DECISION TREES

- When displayed, these series of rules appear as a tree with several branching paths or **splits**.

- The starting point of a decision tree is referred to as the **root** and subsequent branching points are called **nodes**. Nodes that do not split further are then called **leaves**.

- Using our example decision tree:



DECISION TREES

- The structure of a decision tree is determined by what yes/no rules will best predict the outcome variable.

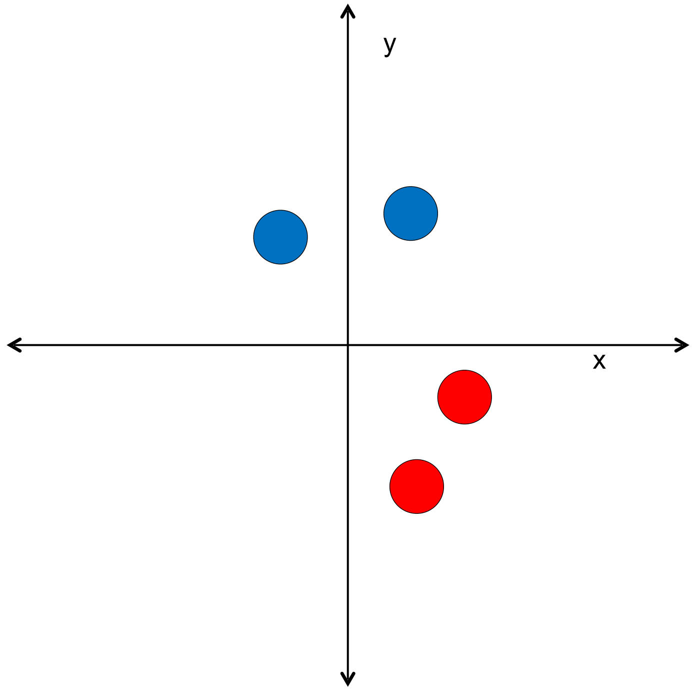
- This is measured at each point of a decision tree by the **gini impurity** which measures the homogeneity of the outcome variable in a dataset from 0 (uniform) to 1 (inconsistent).

- Each rule in a decision tree decreases the gini impurity in the data until it approaches 0.

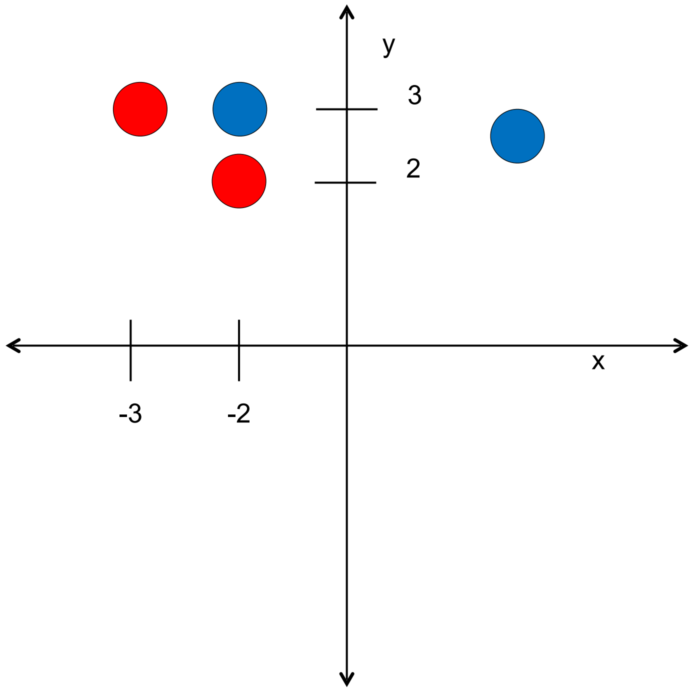
- For regression trees, MSE (or mean squared error) is *often* used in place of gini impurity.

SIMPLE EXAMPLE

Dataset A



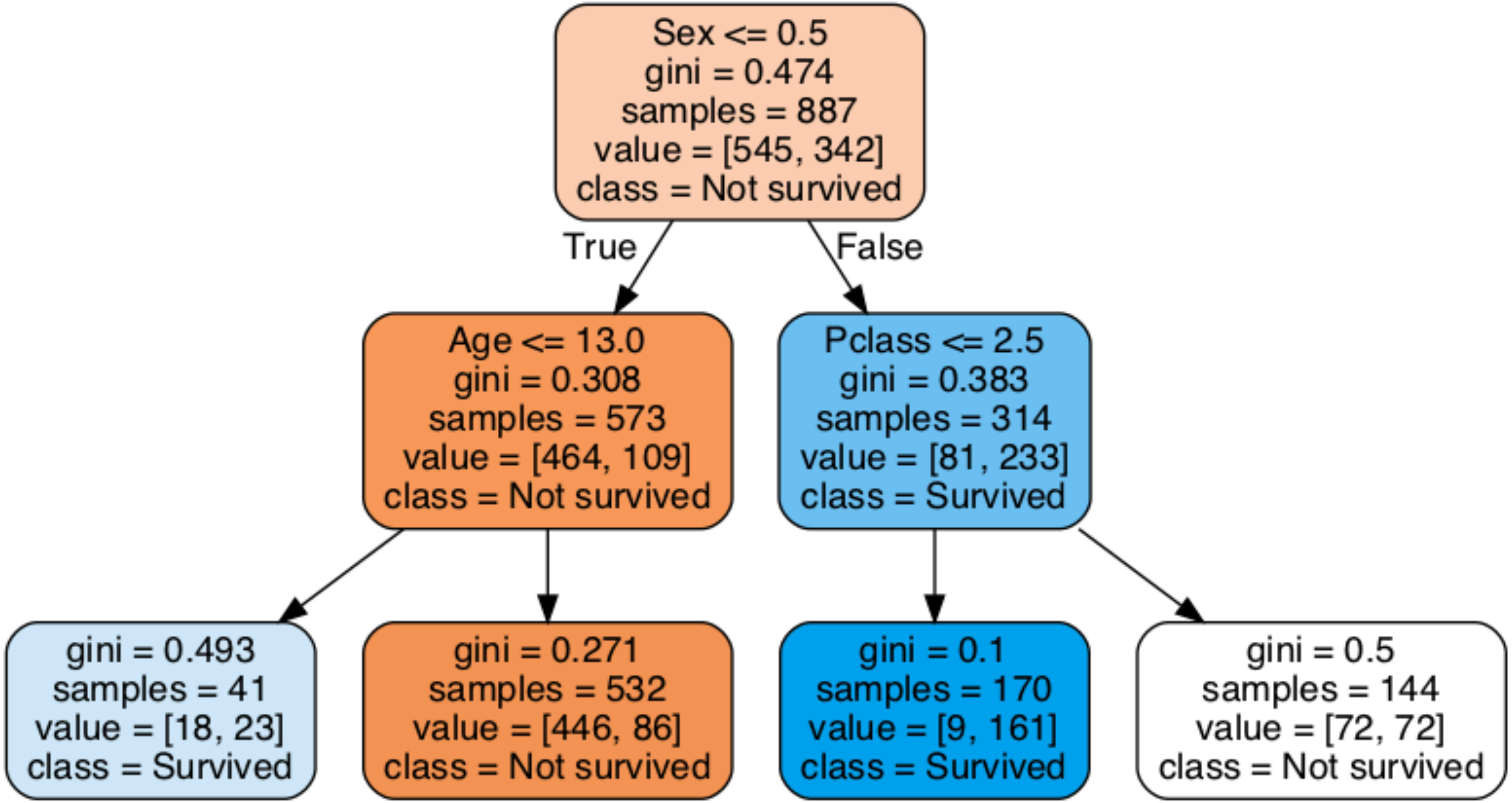
Dataset B



Dataset A – if Y < 0 & X > 0 then RED, else BLUE

Dataset B – if Y > 2 & X > -2 then BLUE, else RED

DECISION TREE OUTPUT



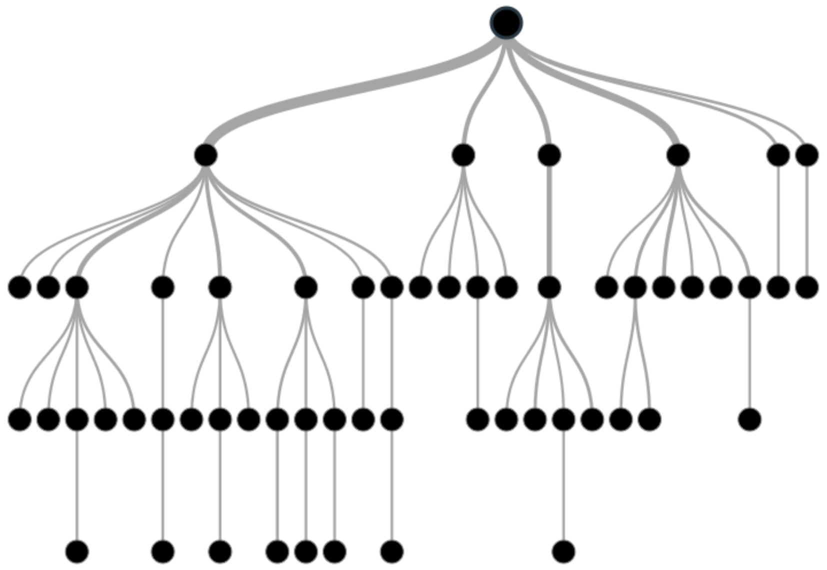
INTRODUCTION

PROS AND CONS OF DECISION TREES

PROS ~~AND CONS~~ OF DECISION TREES

- Decision trees are *non-linear* (a change in a predictor variable has a constant change on the output variable) which gives them more flexibility over linear models (e.g. linear regression).

- Decision trees also produce easily interpreted visuals from which variable importance can be derived.

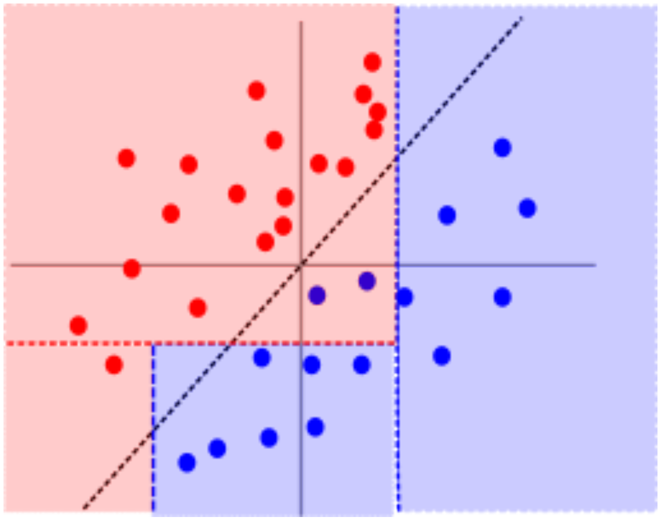


~~PROS AND~~ CONS OF DECISION TREES

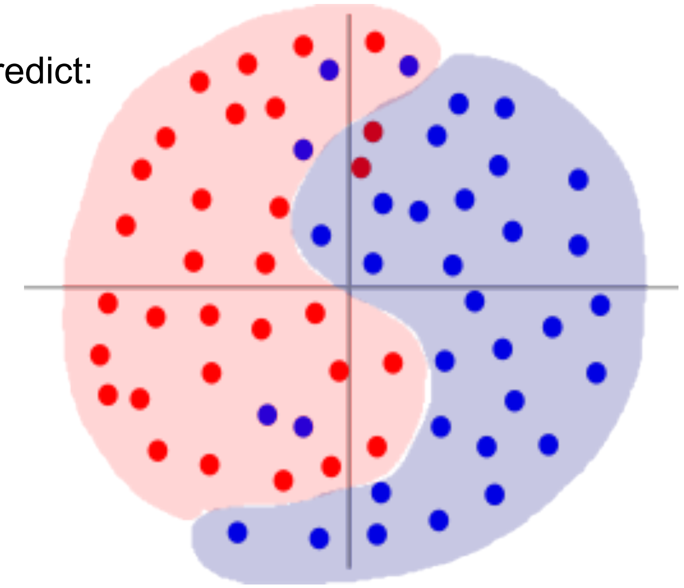
- Decision trees are computationally intensive relative to other models, especially if you don't prune them.

- Decision trees are sometimes too flexible and can easily overfit your data. Cross-validation and tuning are key to keeping decision tree models generalizable.

Fit:

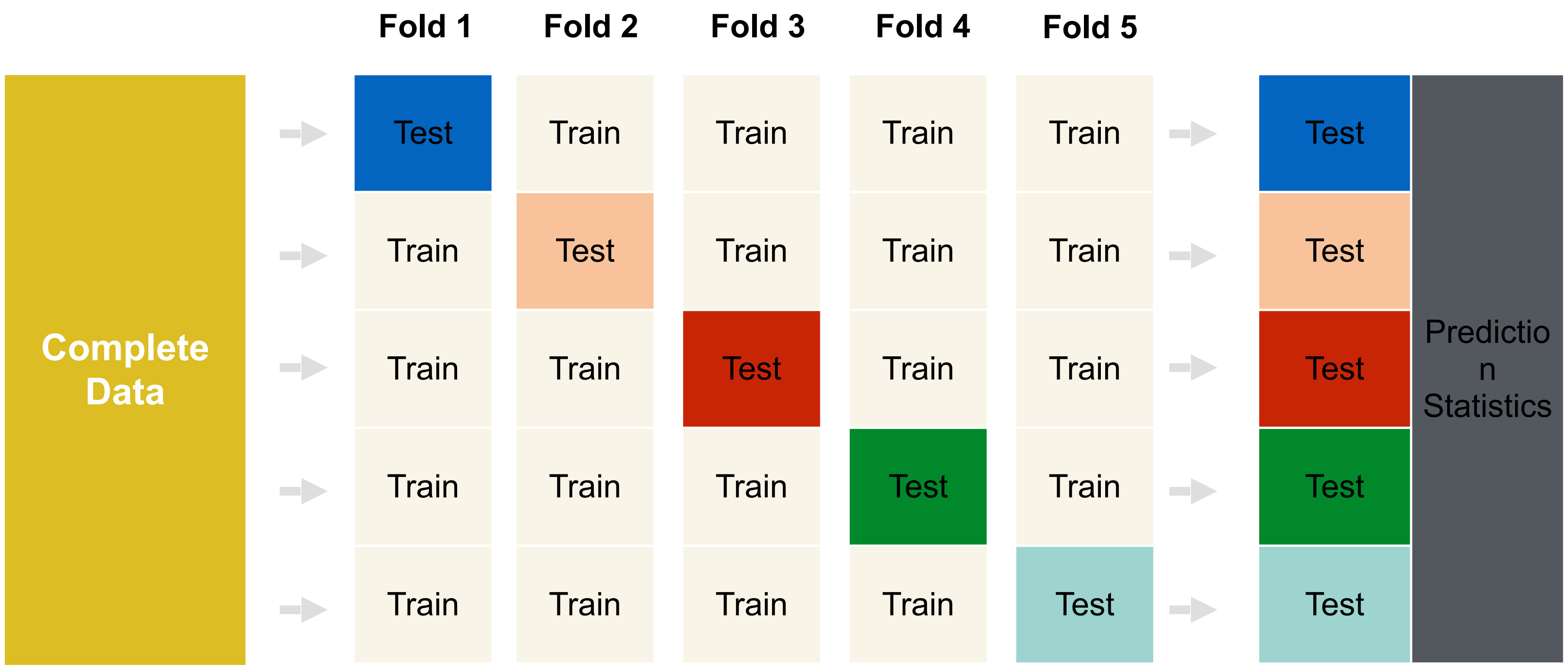


Predict:



K-Fold Cross Validation

Cross validation uses all the data to ensure that model measurements were not biased by a particularly lucky sample



In 5-fold cross validation, one-fifth of the data is used for a testing set in each fold. Then the test set statistics are computed jointly to evaluate how well the model performed.

*Material adapted from “Conversational Analytics” course

INTRODUCTION

ENSEMBLE METHODS

BIAS VS VARIANCE

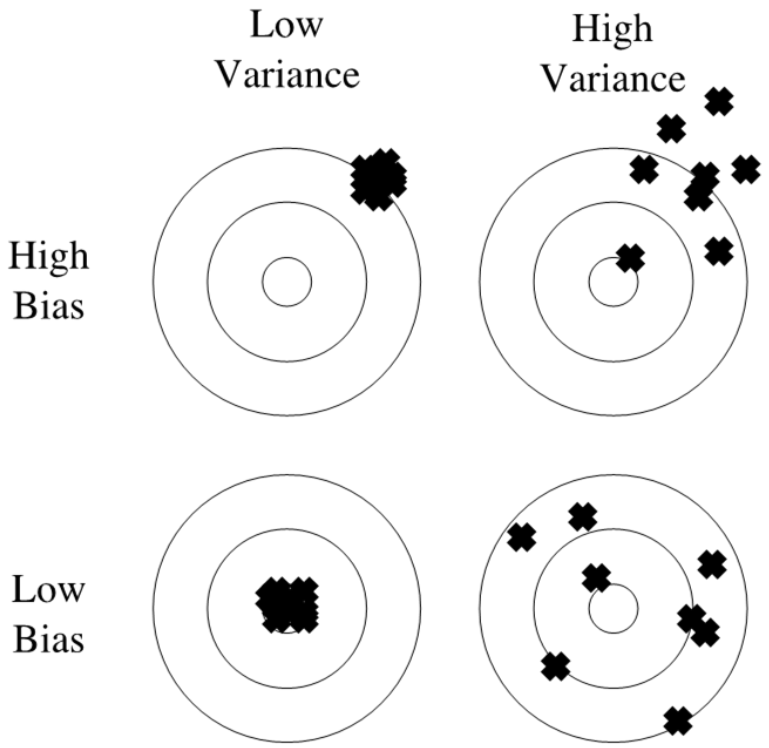


Figure 1: Bias and variance in dart-throwing.

Bias

Definition:

- How close the model's estimate is the to the true value (within our sample)

Variance

Definition:

- How consistent is our predictor over all possible samples drawn from the target population

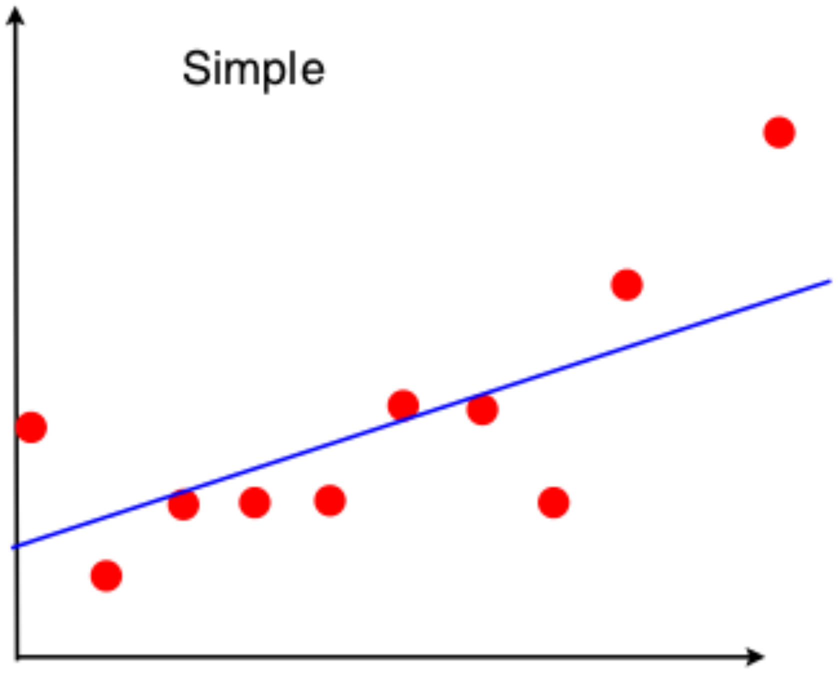
BIAS VS VARIANCE TRADE-OFF

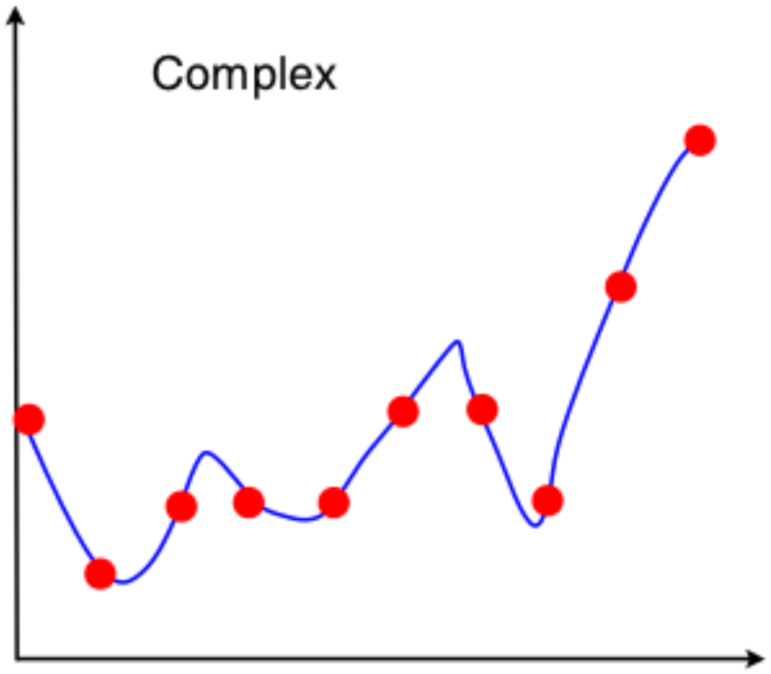
Simple models (few parameters) have higher bias, but lower variance

Result: **Underfitting**

Complex models (many parameters) usually have lower bias, but higher variance.

Result: **Overfitting**





ENSEMBLE LEARNING

Definition:

- **Ensemble methods** are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions[1]

Example:

- Generate 100 different decision trees from the same or different training set and have them **vote on the best classification** for a new example

Key Motivation:

- Reduce the **error rate.** The hope is that it will become much more **unlikely that the ensemble of models will misclassify an example**

ENSEMBLE LEARNING

How much does this cow weigh?

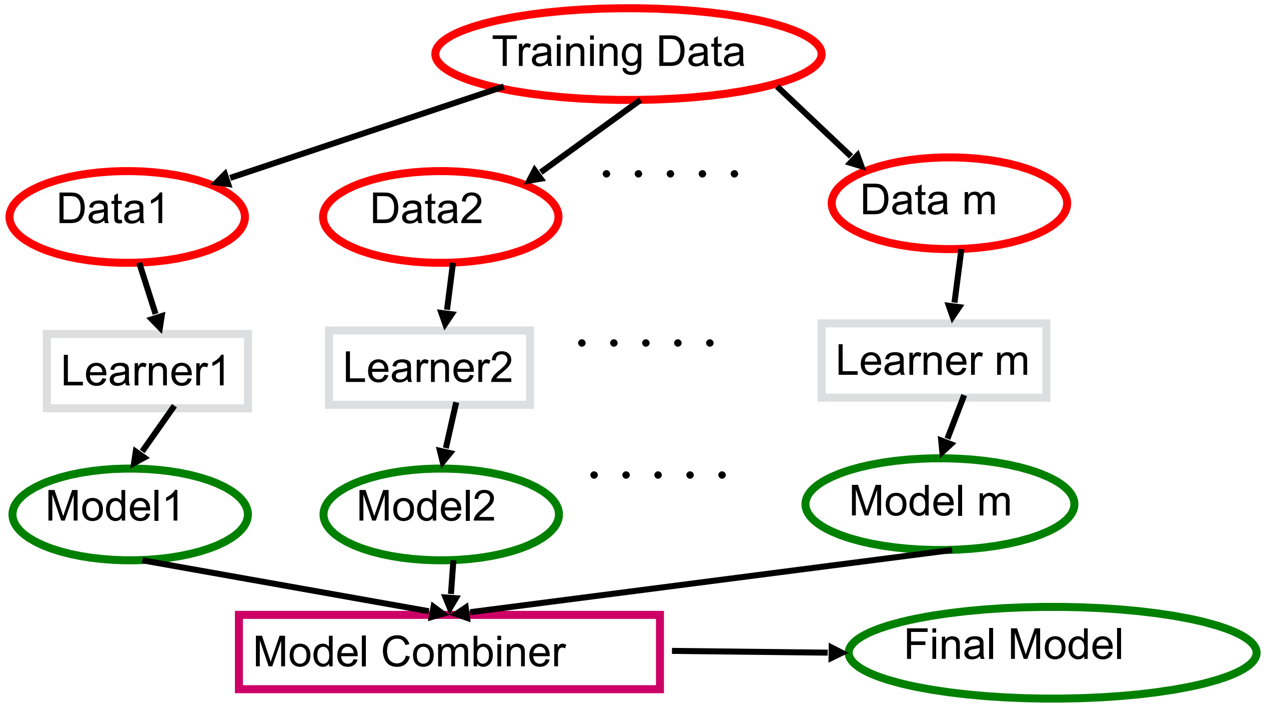
Hypotheses		
1200	1450	1300
900	2000	1100
1500	800	2100
1300	1200	1650
1800	1200	1500
1900	1000	1000
2000	750	600
1500	1250	1350
Average		**1348**
Truth		**1355**

"Wisdom of the Masses"

- First used to explain political concepts
- Theory flawed in practice, but general idea relates well to Ensemble Methods

ENSEMBLE METHODS

Learn multiple alternative definitions of a concept **using different training data** or **different learning algorithms. Combine decisions** of multiple definitions, e.g. using **weighted voting.**



VALUE OF ENSEMBLES

"No Free Lunch" Theorem

- No single algorithm wins all the time
- Model that performs well on one dataset may perform poorly on another

When combing multiple **independent** and **diverse decisions** each of which is **at least more accurate than random guessing**, random errors cancel each other out, **correct decisions are reinforced**

Ensemble Methods Used in Many High-Performing Academic Situations

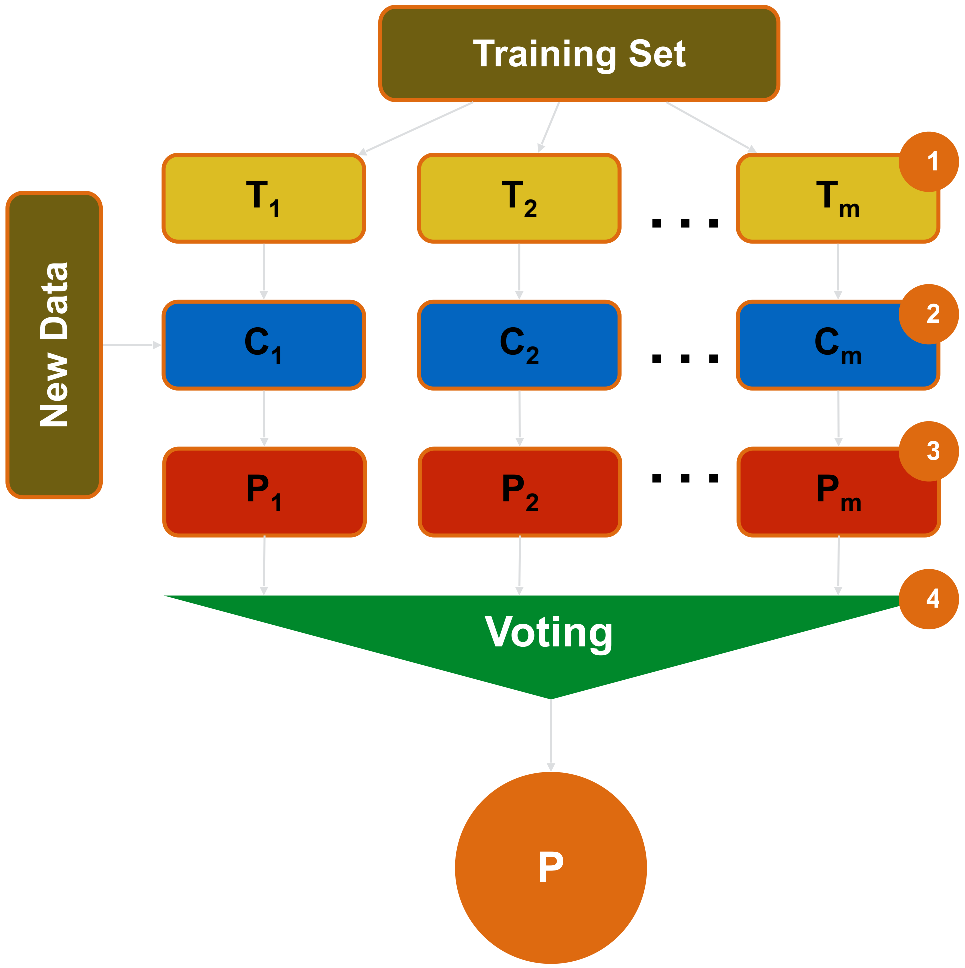
- Netflix Competition Winner
- Numerous Kaggle Competitions
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners

INTRODUCTION

BAGGED TREES

BAGGING DECISION TREES

Voting ensemble built from bootstrapped training samples



1 Bootstrap samples

- Draw multiple bootstrapped datasets (sample one point at a time with replacement) from our original training dataset
- The process creates m datasets (T_1,T_2,…,T_m) of equal length to the original

2 Classification Models

- Build a series of classifiers (C_1,C_2,…,C_m) from the samples above
- No requirement on the type of classifiers used (but usually trees)

3 Predictions

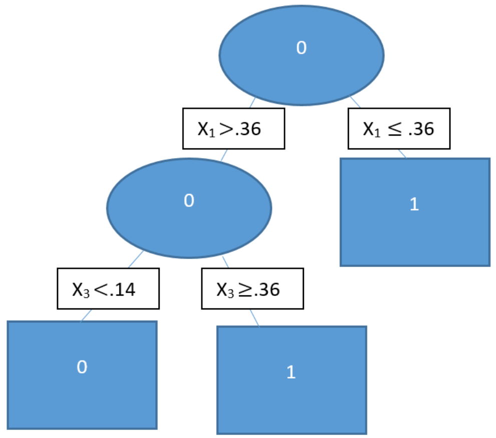
- As new data comes in it gets scored by each model

4 Final Prediction

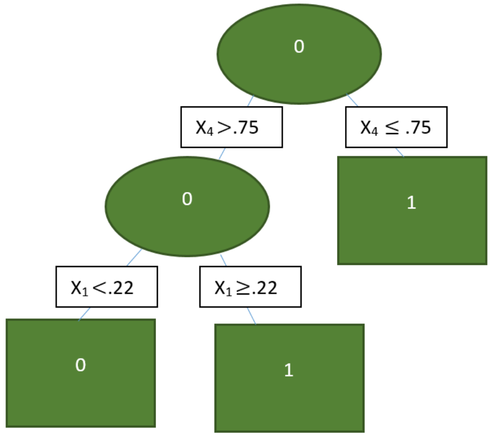
- The final prediction is based on a vote (for classification) or averaging (for a regression) among the various individual predictions
- The degree of variance reduction due to the bootstrap depends on the type of classifier chosen

BAGGING DECISION TREES

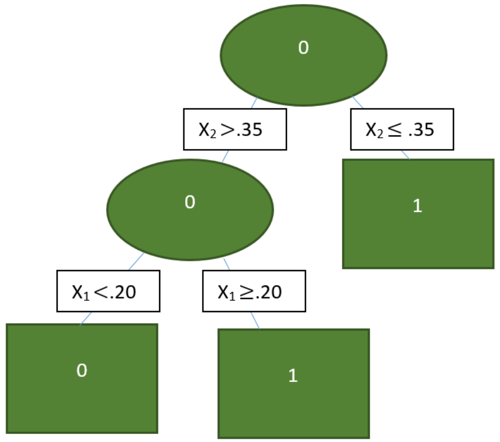
Original Tree



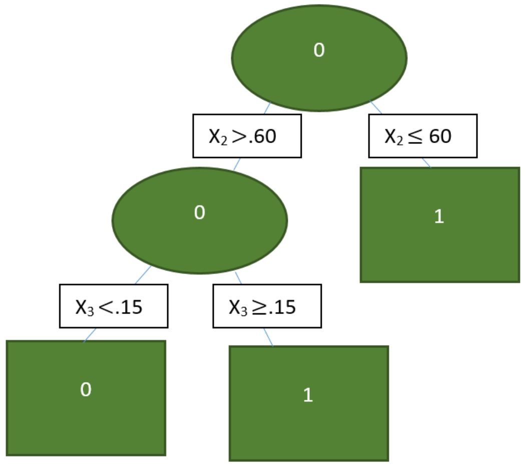
Bootstrap Tree 1



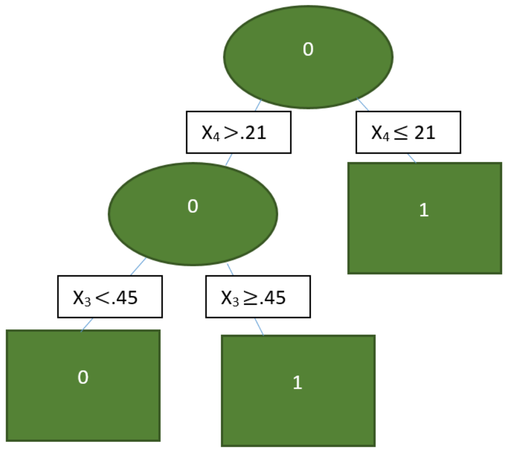
Bootstrap Tree 2



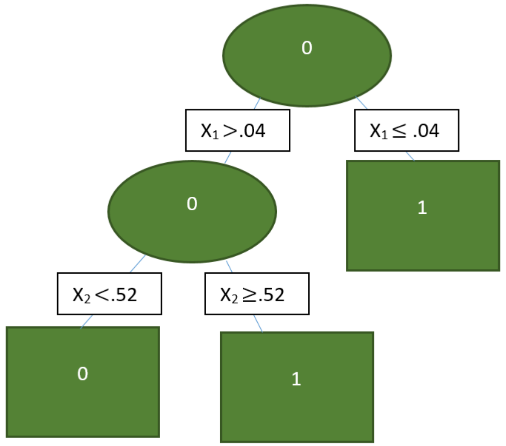
Bootstrap Tree 3



Bootstrap Tree 4



Bootstrap Tree 5



PROS AND CONS OF BAGGING DECISION TREES

Advantages

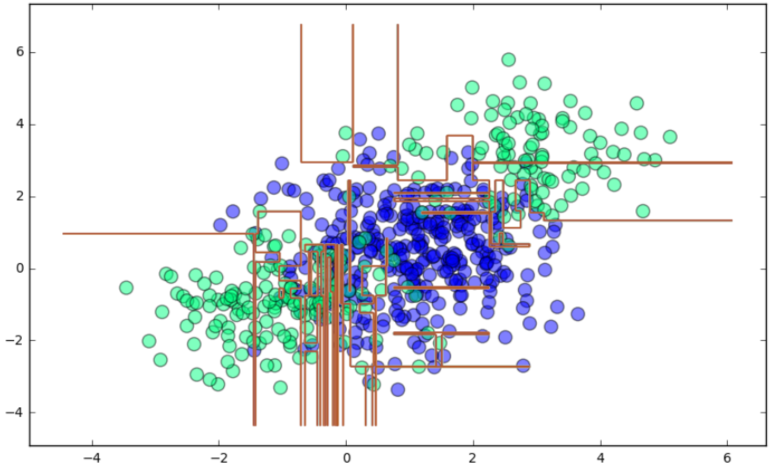
- **Reduces variance** in comparison to regular decision trees
- Can provide **variable importance measures**
- Can easily handle **qualitative (categorical) features**
- Out of bag (OOB) estimates can be used for **model validation**

Disadvantages

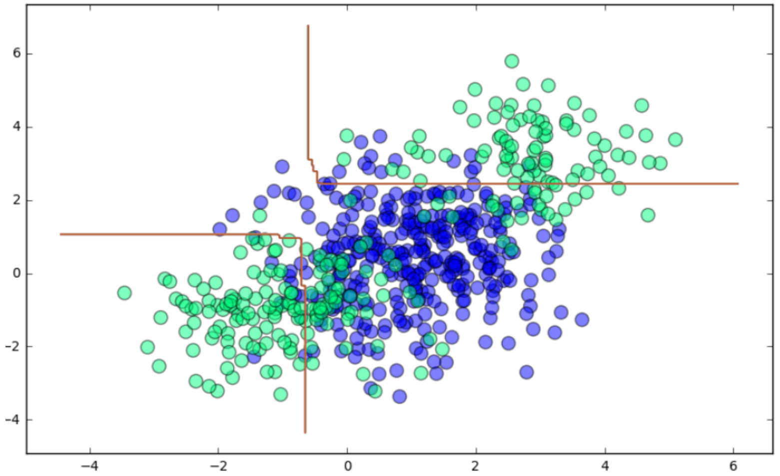
- Bagged models can be **more difficult to interpret** than single decision trees
- When **multiple predictors** are used, it is not always clear **which ones are important** in a bagged model

BAGGING DECISION TREES

Single Decision Tree
Training Accuracy: 100%



Bagging
Training Accuracy: 82.3%



We lose training accuracy with the bagging algorithm, but bagging **generalizes to the data** and may perform better on new observations than the single decision tree due to a **smoother decision boundary.**

INTRODUCTION

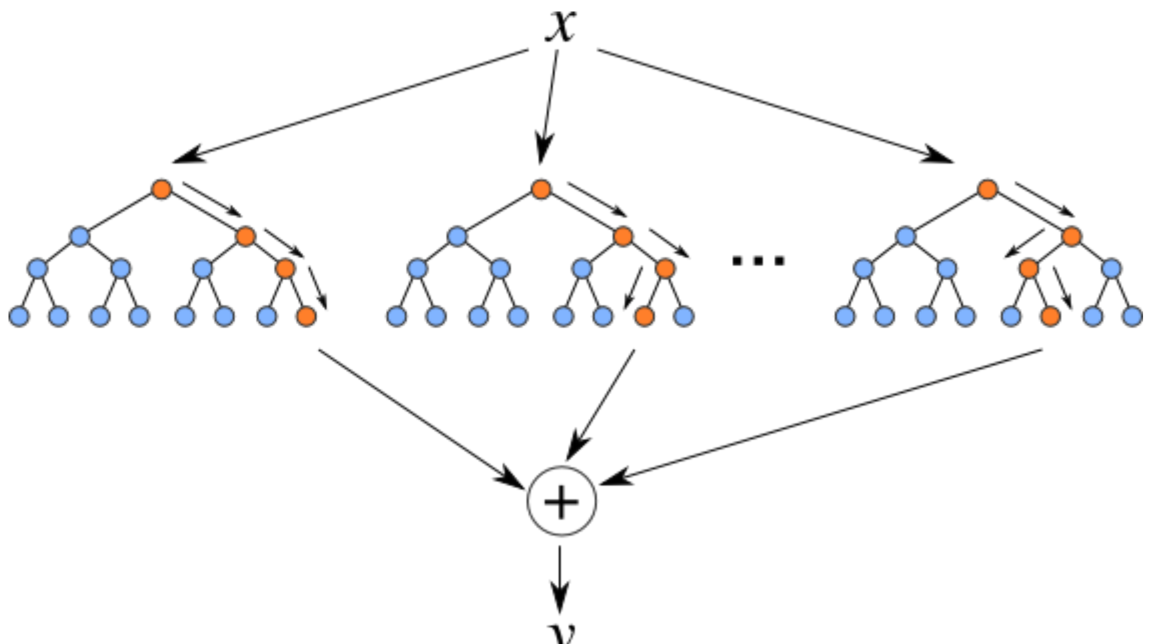
RANDOM FORESTS

BAGGING VS. RANDOM FORESTS

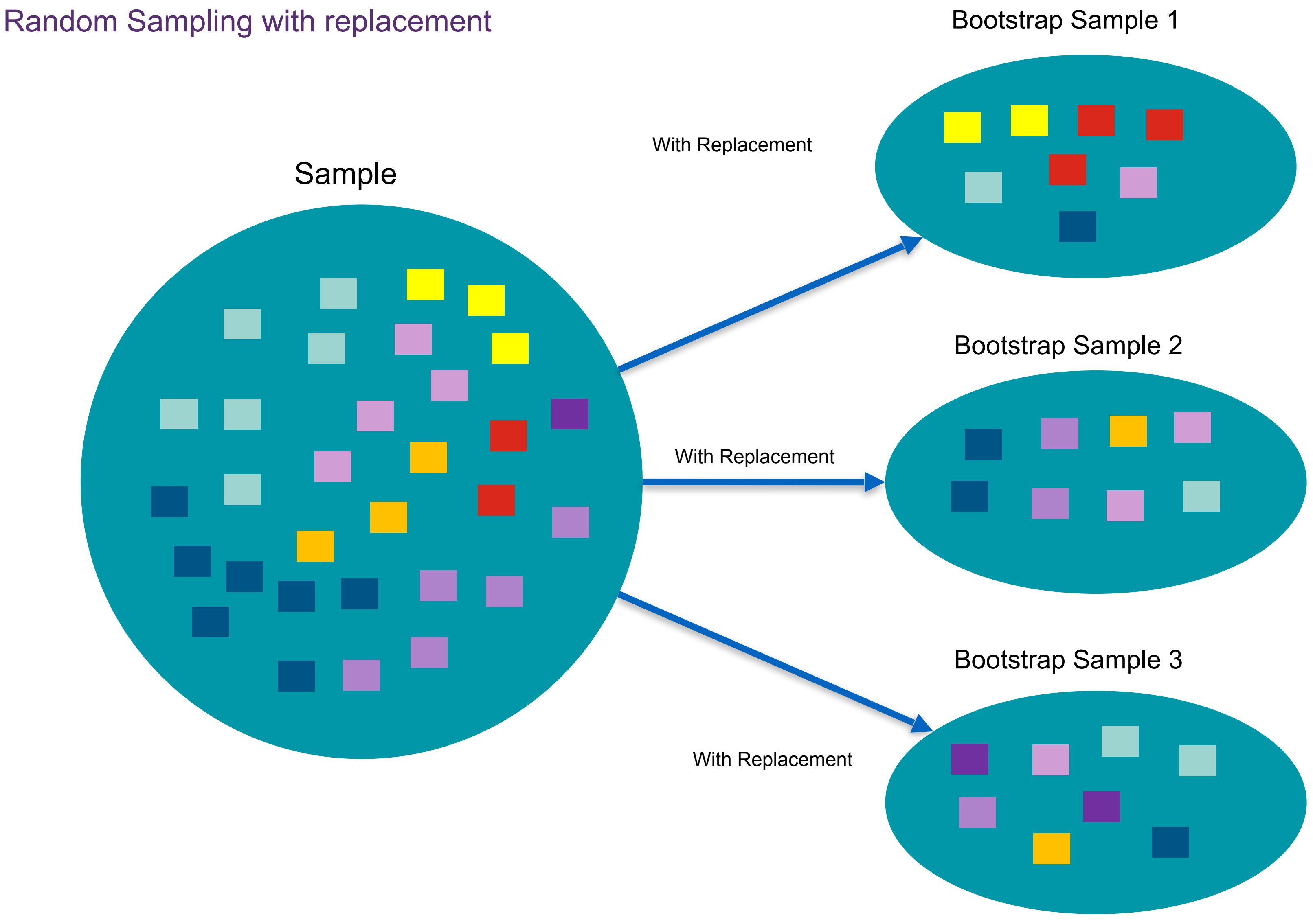
- Suppose we have a set of *m* predictors. Further, suppose **one** of these predictors is a **strong predictor** for the outcome and **many** of the predictors are **moderately strong**
- In a collection of bagged trees, **most or all of the trees will use the strong predictor for the top split**. As a result, the bagged trees will all be similar and their **predictions will be correlated**
- ***Will averaging a number of highly correlated trees result in a significant reduction in variance over a single tree? Answer: NO.***

RUNNING THROUGH THE RANDOM FORESTS

- Random forest models are one of the most widespread classifiers used because they are relatively simple to use and avoid overfitting.
- They do this by **ensembling** or aggregating the results of several individual decision trees.

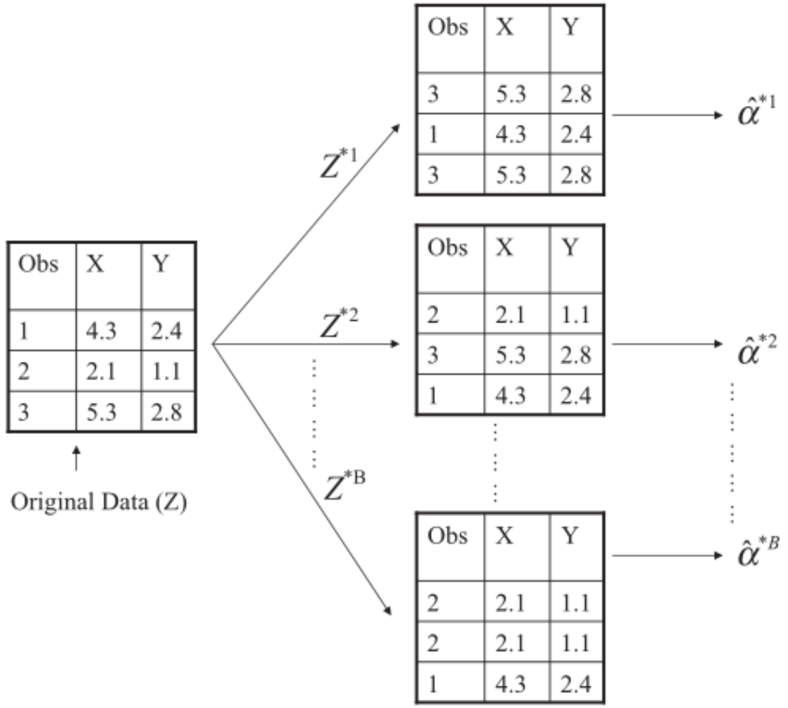


BOOTSTRAPPING



RUNNING THROUGH THE RANDOM FORESTS

- Random forests generates many decision trees using another resampling method – **bootstrapping**.



- Bootstrapping differs from cross-validation samples and allows replacement.

RUNNING THROUGH THE RANDOM FORESTS

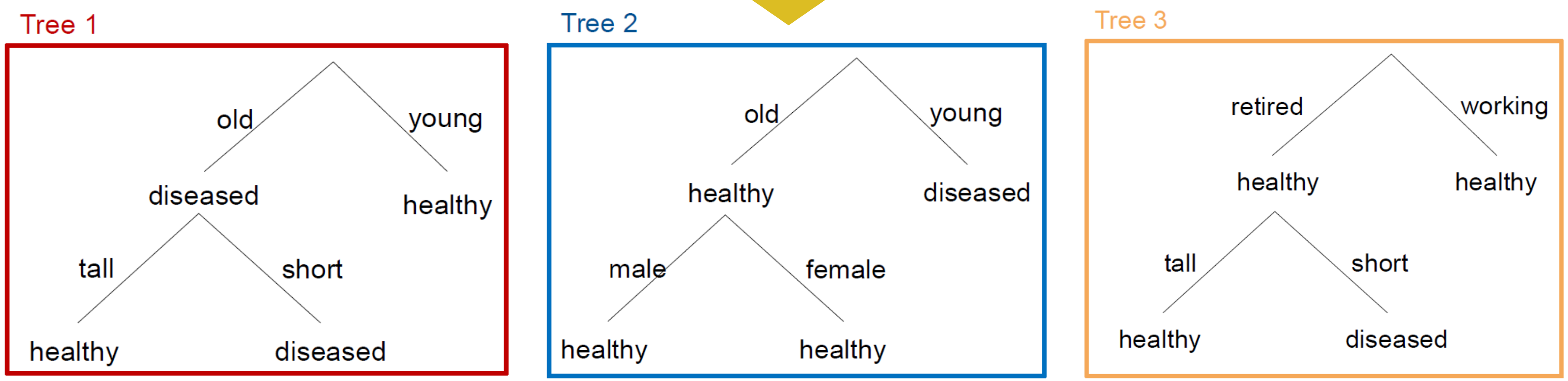
- For every bootstrapped sample, a decision tree is built and then the results are aggregated to form a random forest.

- The idea is that individual trees are likely to overfit, but a set of trees generated from random samples of the original data are unlikely to overfit because each sample will be different.

- *Only the most significant decision rules will be the same across different trees in the same forest.*

RANDOM FOREST: EXAMPLE

Inputs:
young, working, male, tall

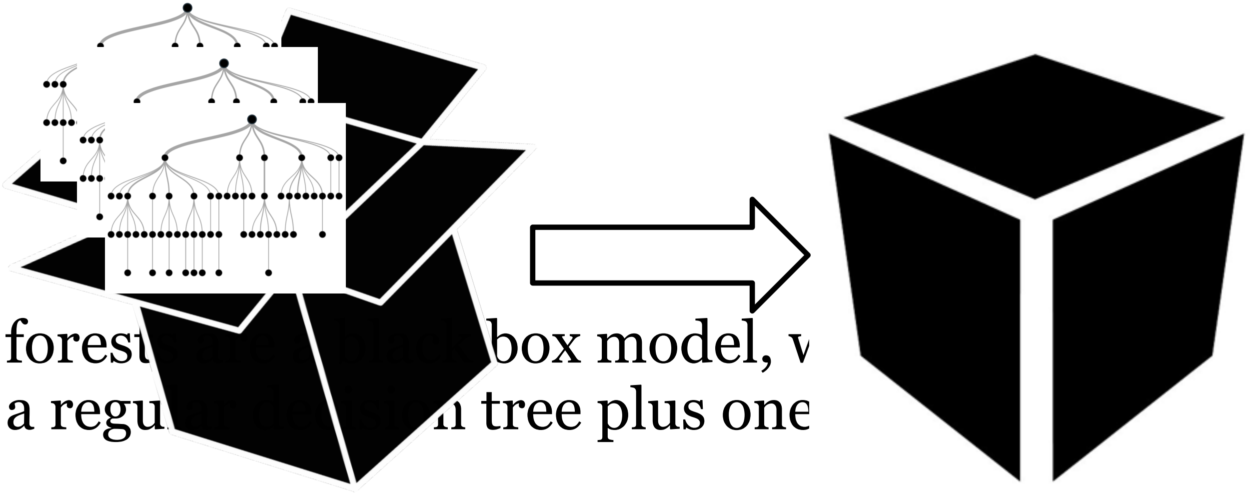


Outputs:
healthy, diseased, healthy

Majority Vote:
Healthy

CONS OF RANDOM FORESTS

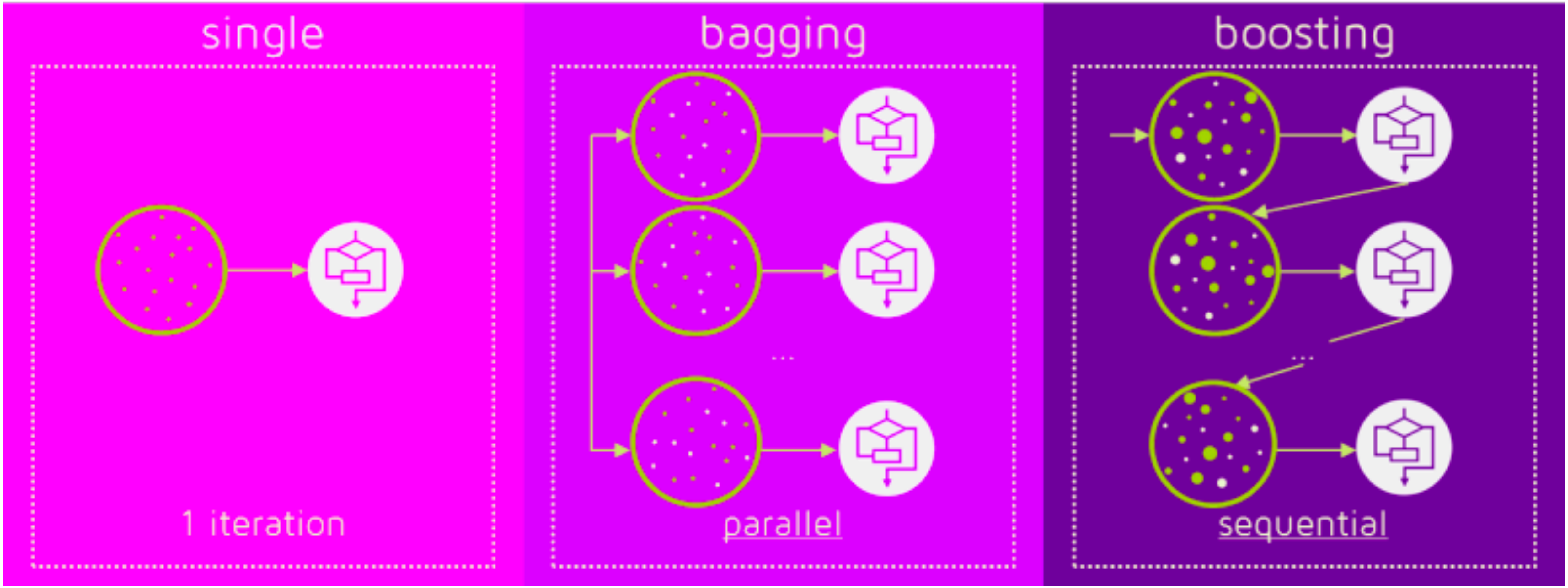
- This comes with a major tradeoff – random forests are a *black box model* so we lose the interpretability and visualization of decision trees.



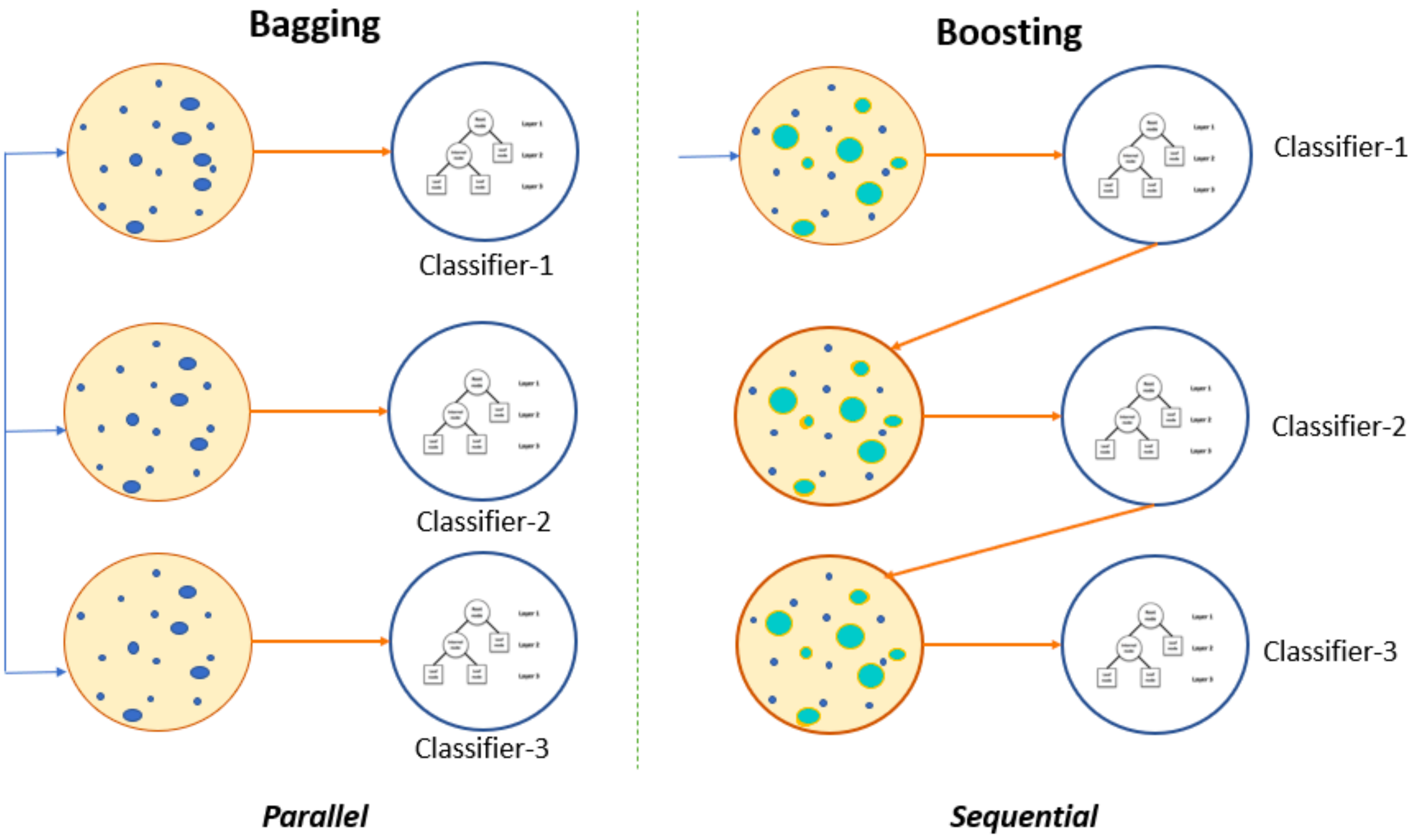
- Even though random forest … box model, w… s to all of the same tuning parameters as a regu… n tree plus one … per of trees to build before ensembling.

BOOSTING

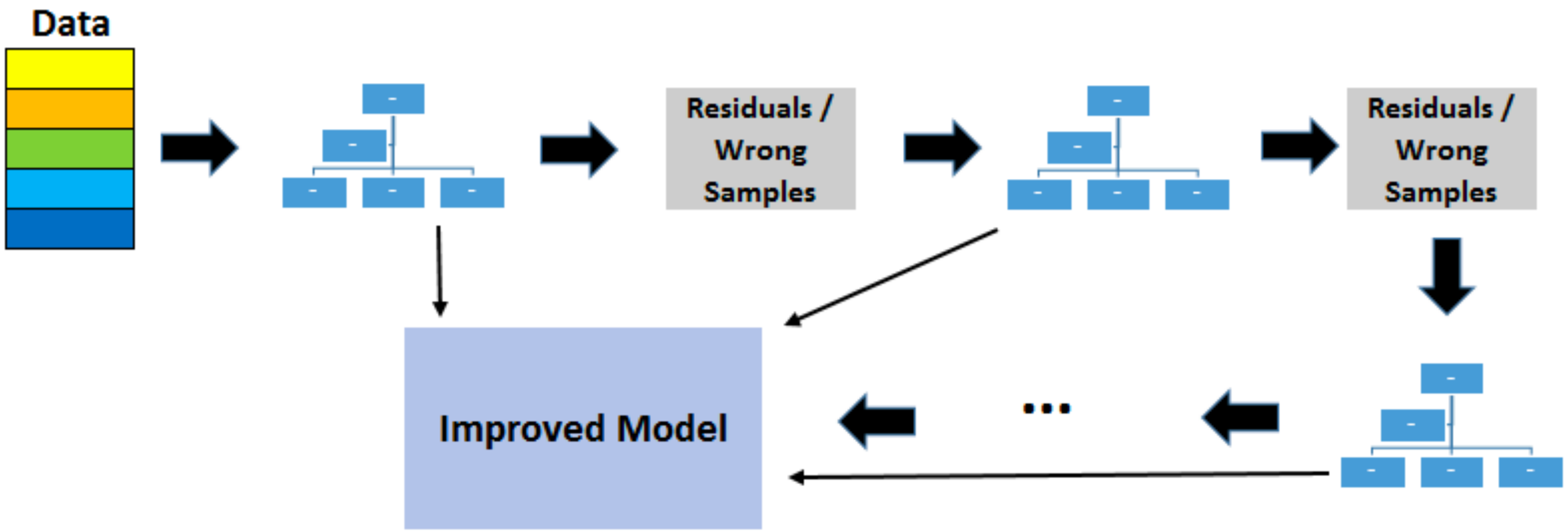
Bagging vs. Boosting



Bagging vs. Boosting



BOOSTING



Data
Residuals /
Wrong
Samples
Residuals /
Wrong
Samples
Improved Model
...

SUMMARY

SUMMARY

Decision Trees

- ✓ Yield insight into decision rules
- ✓ Computationally efficient/ fast
- ✓ Easy to tune parameters
- ✗ High variance in predictions and overfitting

Bagging

- ✓ Easy to tune parameters
- ✓ Smaller prediction variance (sometimes)
- ✗ Difficult to interpret decision rules- often viewed as “black box”
- ✗ May not reduce variance if features are correlated

Random Forest

- ✓ Smaller prediction variance, even with correlated predictors
- ✓ Easy to tune parameters
- ✗ Difficult to interpret decision rules- often viewed as “black box”

CITATIONS

CITATIONS

- *Decision Tree Visualization:* https://littleml.files.wordpress.com/2012/01/screen-shot-2012-01-23-at-10-00-17-am1.png
- *90's Flowchart*, Munroe, Randall: https://xkcd.com/210/
- *Questions on some data-mining algorithms*: https://stackoverflow.com/questions/4084668/questions-on-some-data-mining-algorithms

CITATIONS

- *An Introduction to Statistical Learning*, James, G et al (2013): http://www-bcf.usc.edu/~gareth/ISL/getbook.html
- *The Lorax (Character)*, Seuss Wikia: http://seuss.wikia.com/wiki/The_Lorax_(Character)
- *Classification and Regression Trees*, Cosma Shalizi: http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf