Forest Service
U.S. DEPARTMENT OF AGRICULTURE

Soil Mapping and Classification in Google Earth Engine

Day 2:
Random Forests

Juliette Bateman (she/her)
Remote Sensing Specialist/Trainer,
juliette.bateman@usda.gov

Lila Leatherman (they/them)
Remote Sensing Specialist/Trainer
lila.leatherman@usda.gov

Geospatial Technology and Applications Center | GTAC
USDA Forest Service

Mapping Our Future Together

Housekeeping

- **Keep video off and stay on mute**
- **When you have questions:**
  - Raise hand in Teams
  - Respond in chat box
  - Q + A at the end
- **Closed captions are available**
- **Take care of your body!**

Remember to record!

Day 2 Agenda

- **Afternoon**
  - 13:45-14:45 – Presentation: Intro to Random Forests
  - 14:45-15:00 – Demo: (Ex 4.2) Run a Random Forest Regression

  15:00-15:05 – Break
  - 15:05-15:30 – Presentation Accuracy Assessment

Learning objectives

- Understand how Random Forests is distinct from classification and regression trees
- Understand the difference between classification and regression trees
- Learn key parameters and considerations for employing Random Forests

Random Forests

- **What:** sophisticated ensemble machine learning algorithm
- **Who:** developed by Leo Breiman and Adele Cutler
- **When:** 2001
- **Why:** need to correct for decision trees overfitting training data
- **How:** …we'll get to this in a bit

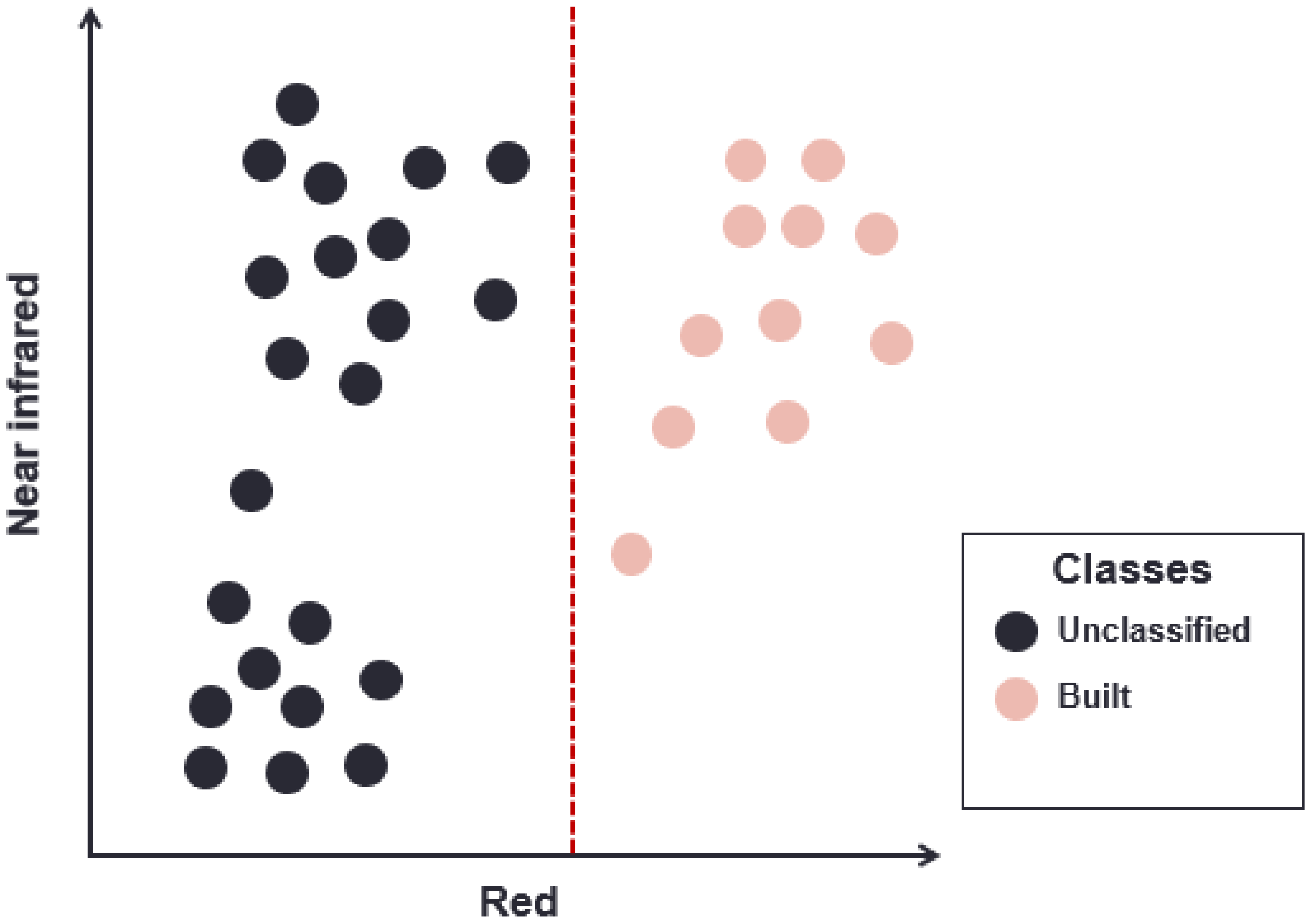
Random Forests

- **What**
  - Sophisticated data mining tool
  - Ensemble of decision trees
  - Few parameters to set (easy to use for the layman)
  - Underlying distribution of data irrelevant (parametric and non-parametric distributions are accepted)
  - Not sensitive to bias or effects of high variance

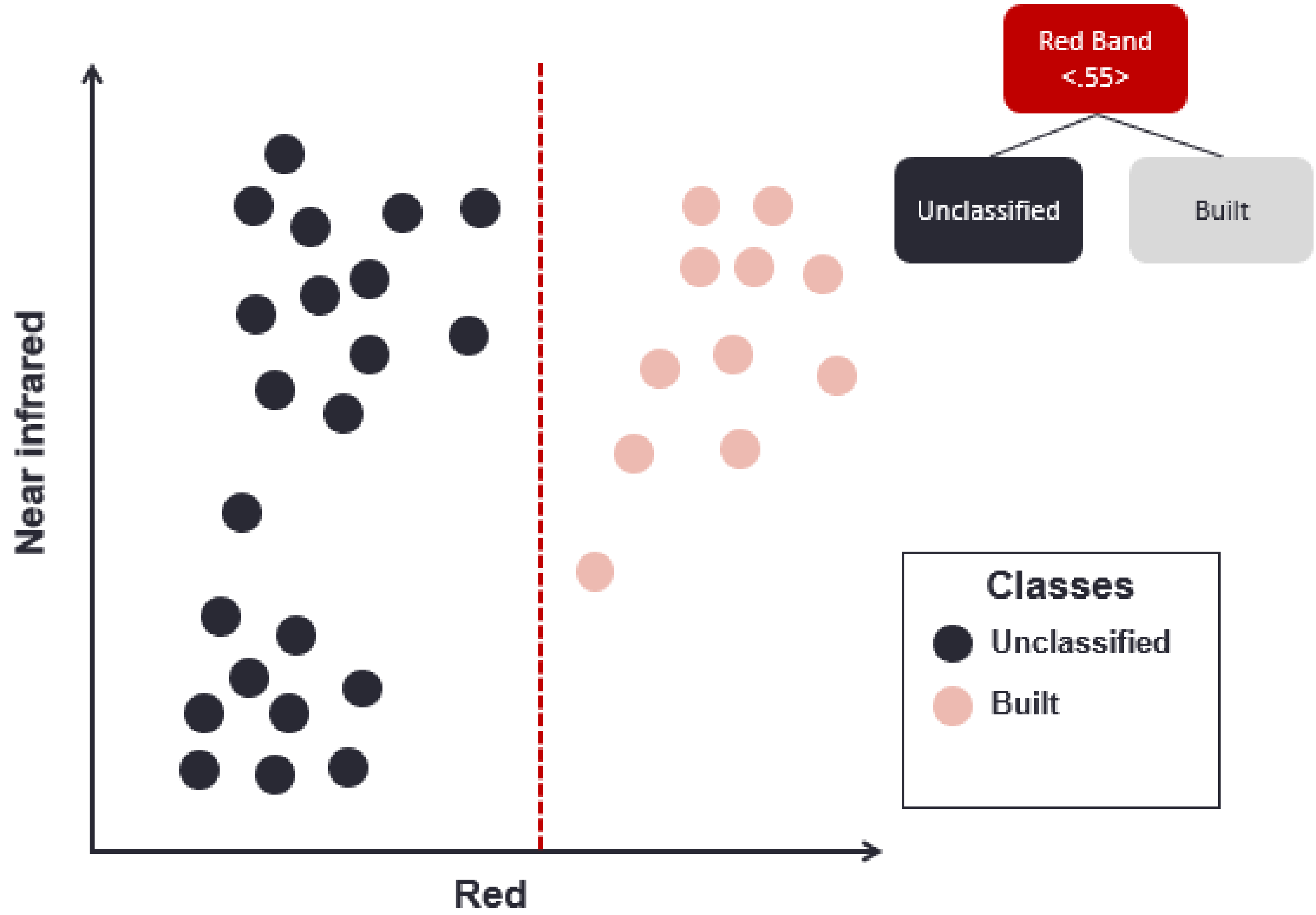
Classification and Regression Trees

- **RF is based in CART method**
- **How CART works:**
  - CART seeks the most ideal splitting point and chooses the variable with the highest discriminating power
  - Uses an impurity function to test splitting thresholds
  - Recursive binary partitioning
    - Recursive (over and over), binary (yes/no questions/criteria), partitioning (splitting the data)

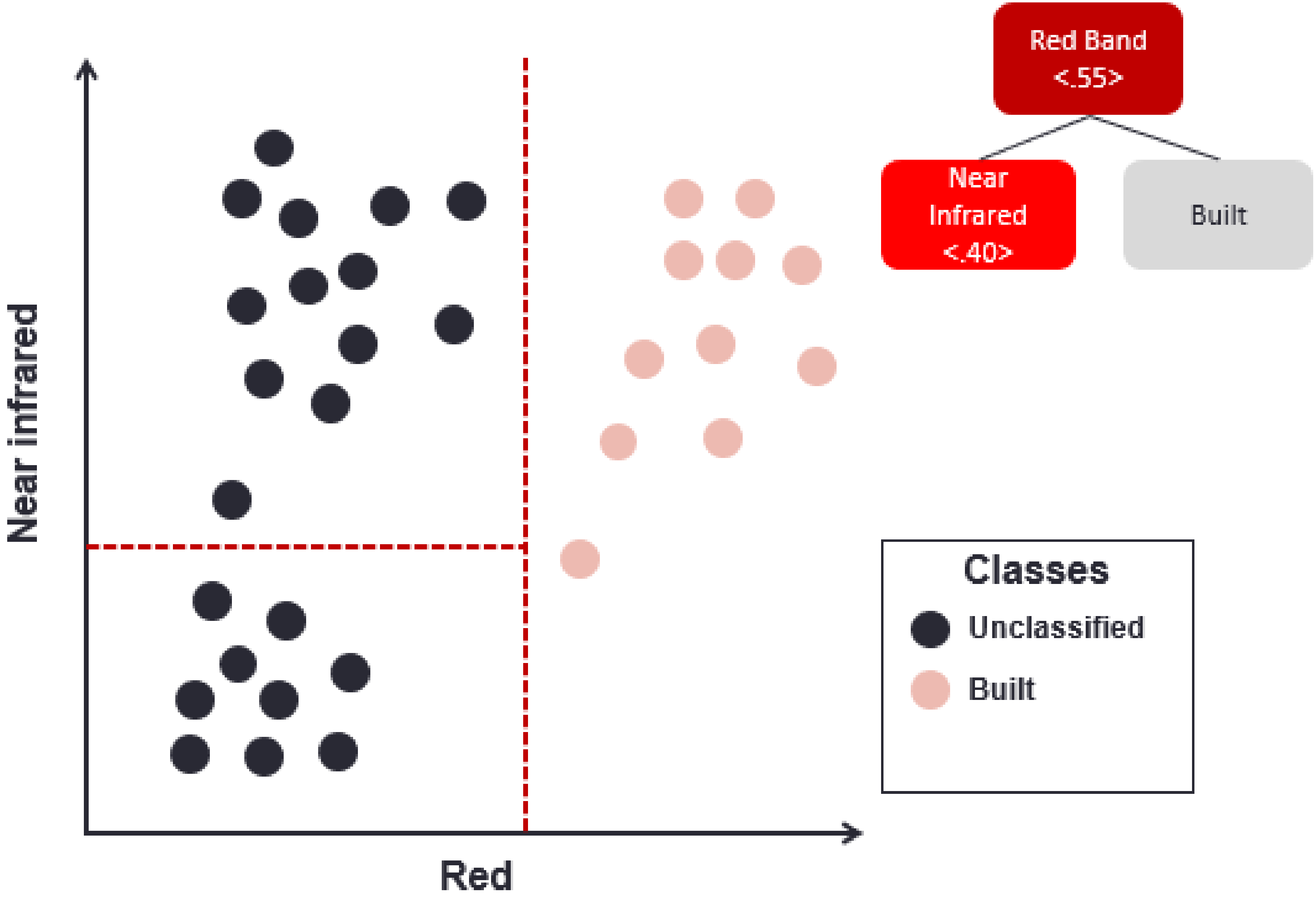
CART Feature Space



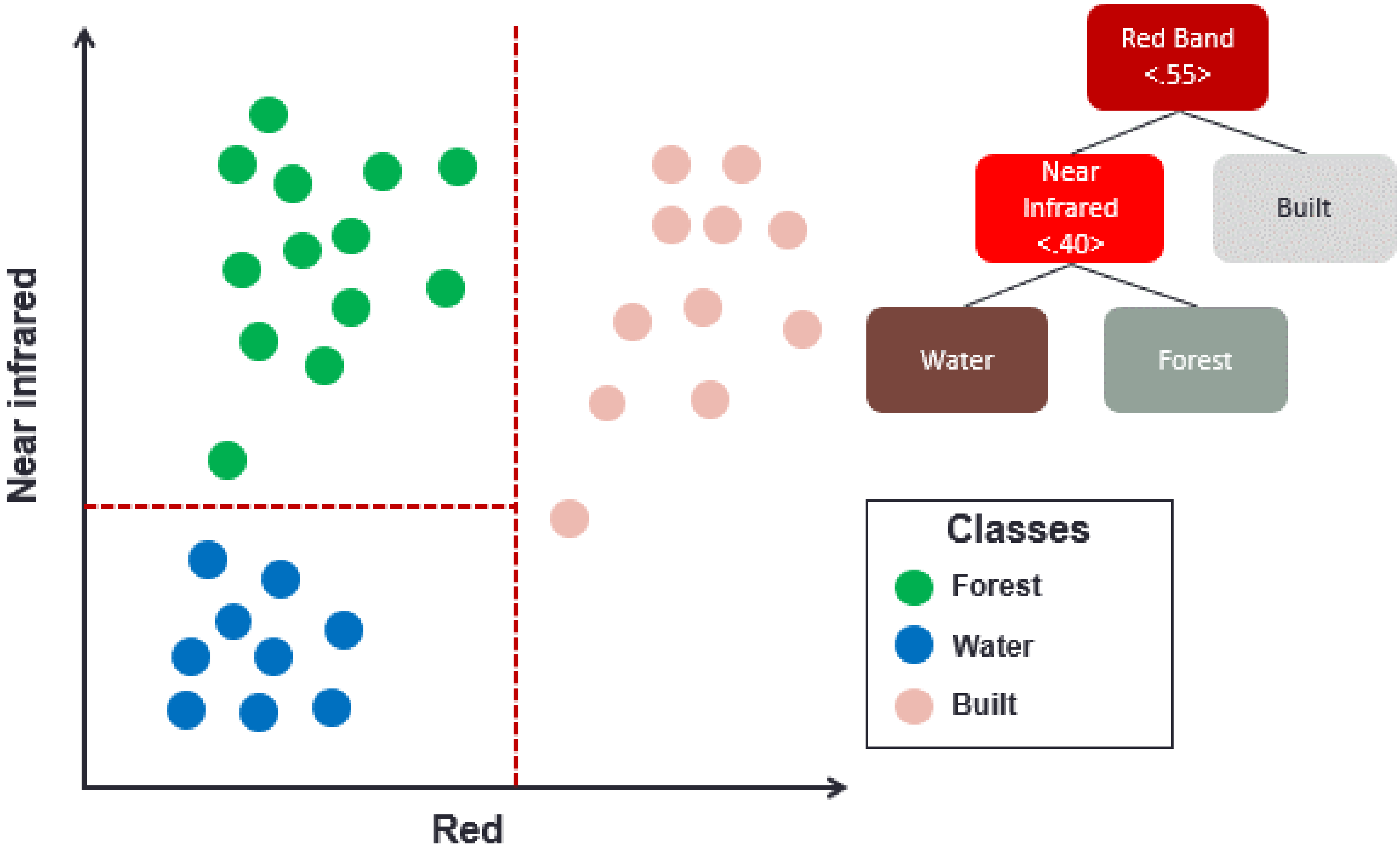
CART Feature Space



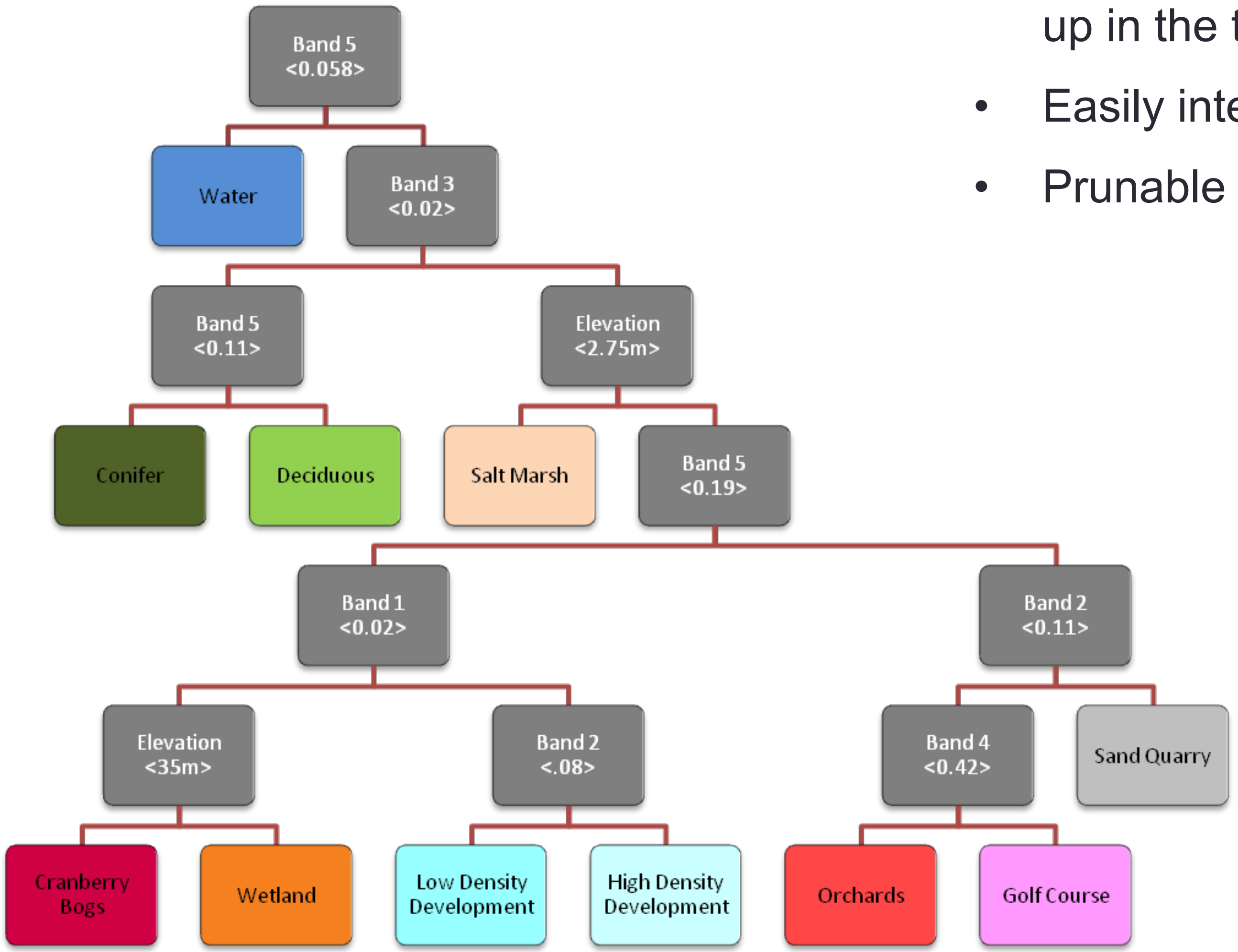
CART Feature Space



CART Feature Space



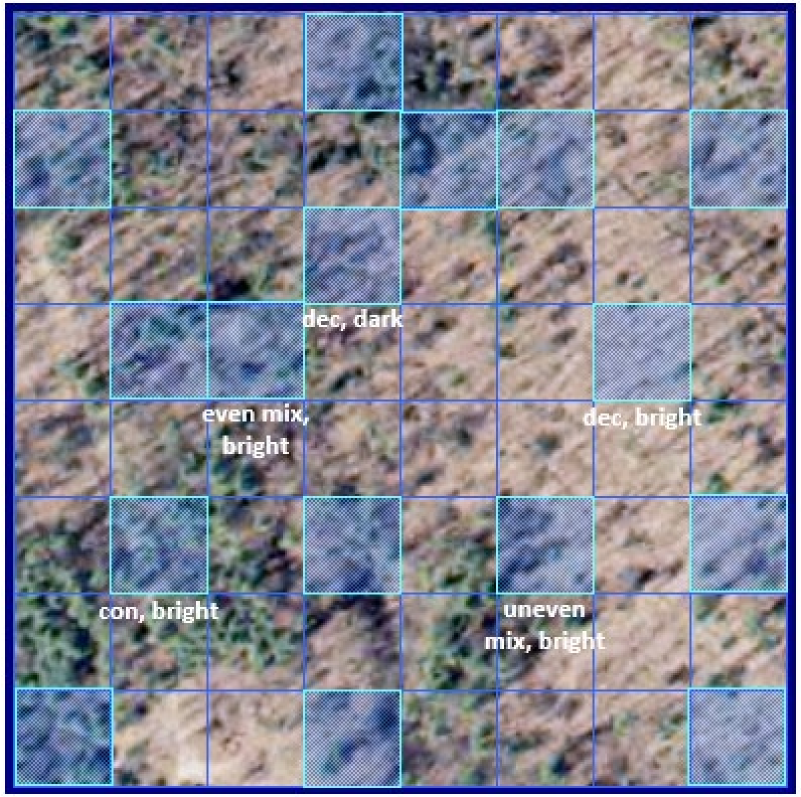
Classification tree example

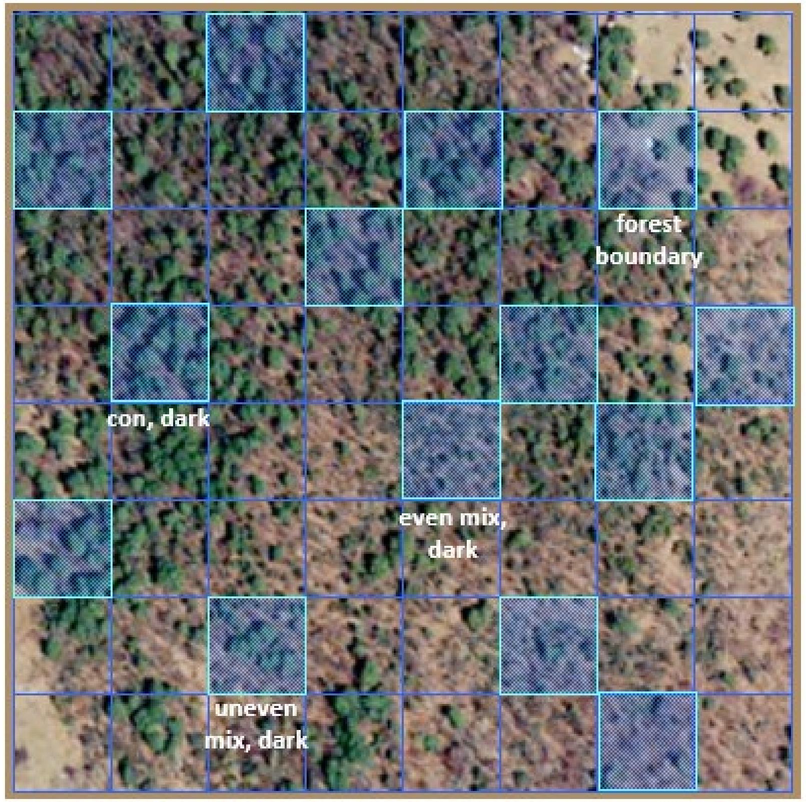


- More informative splits higher up in the tree
- Easily interpretable
- Prunable

High within-class variability

Widely variable sub-pixel mixing effects associated with moderate resolution data

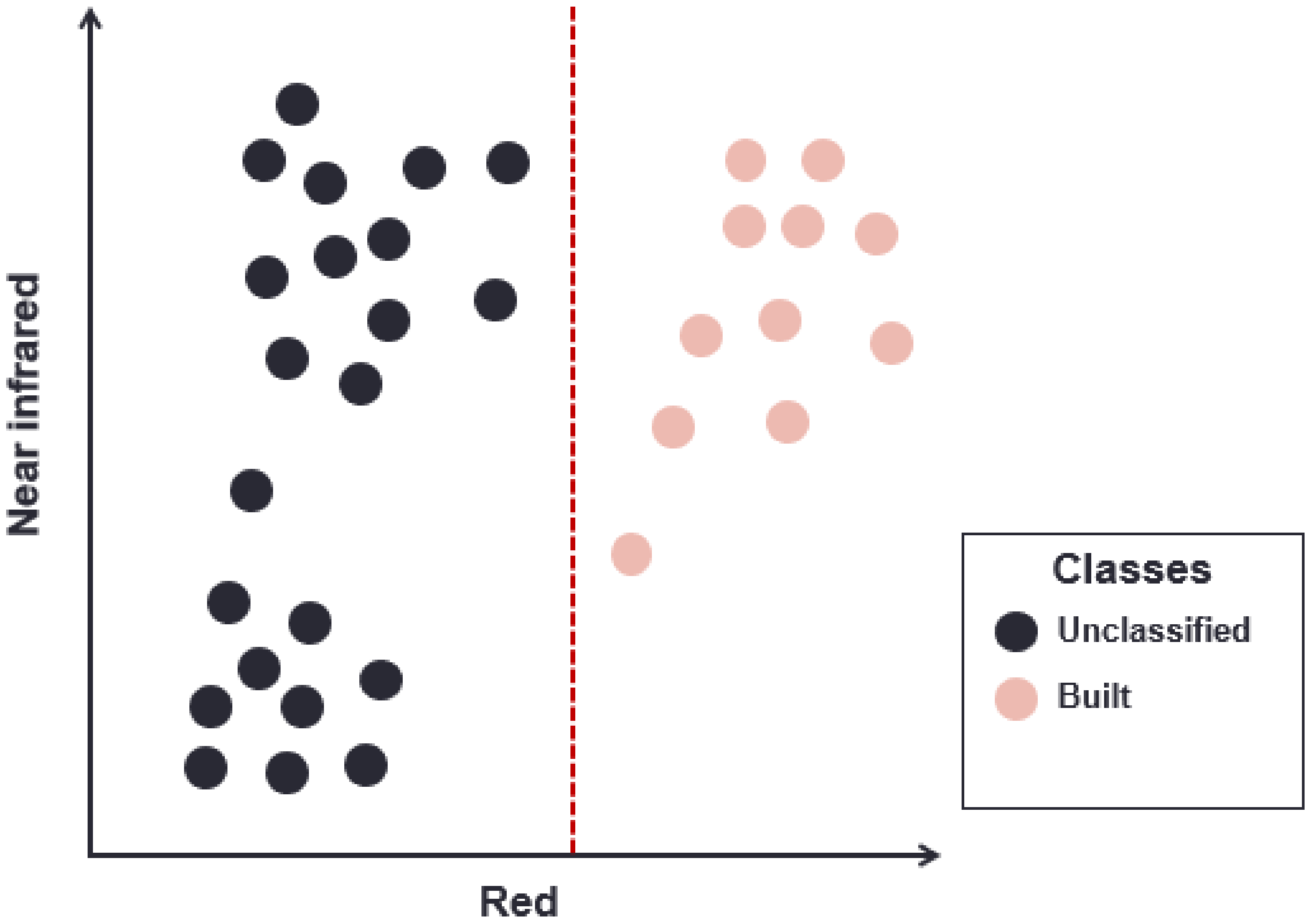




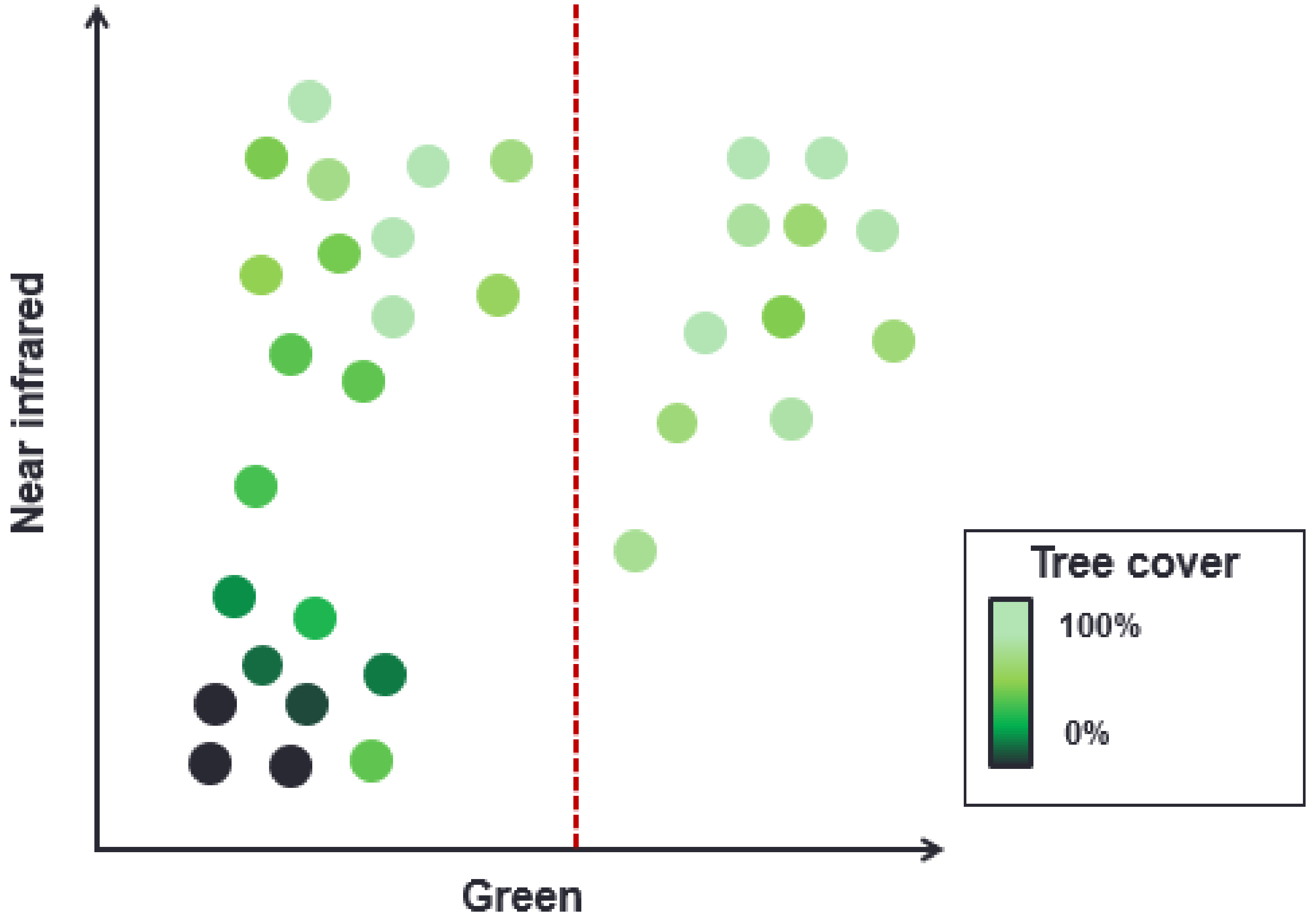
Regression

- Works similarly to classification – but assigns continuous values to end “leaves,” rather than categorical bins

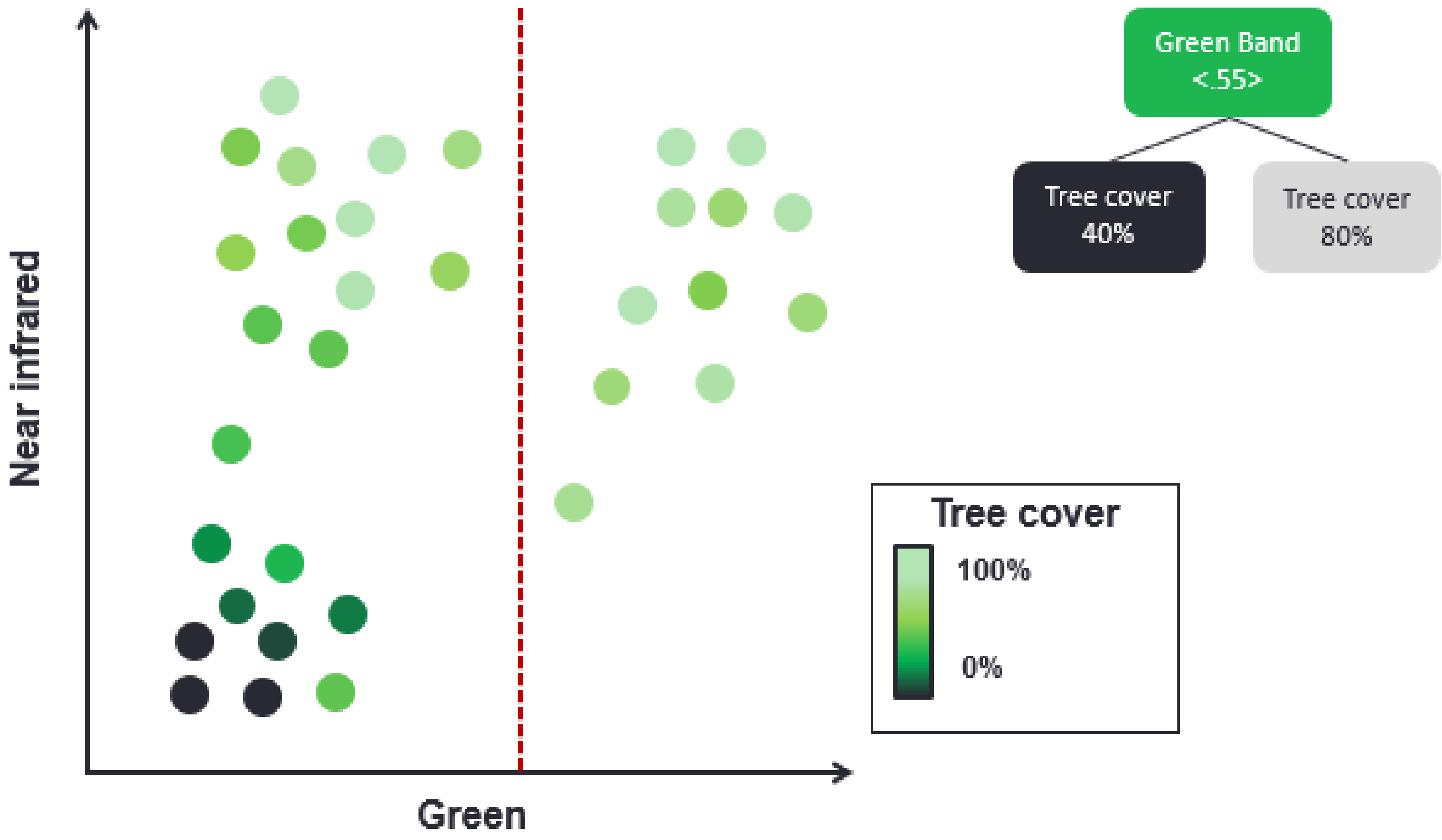
CART Feature Space



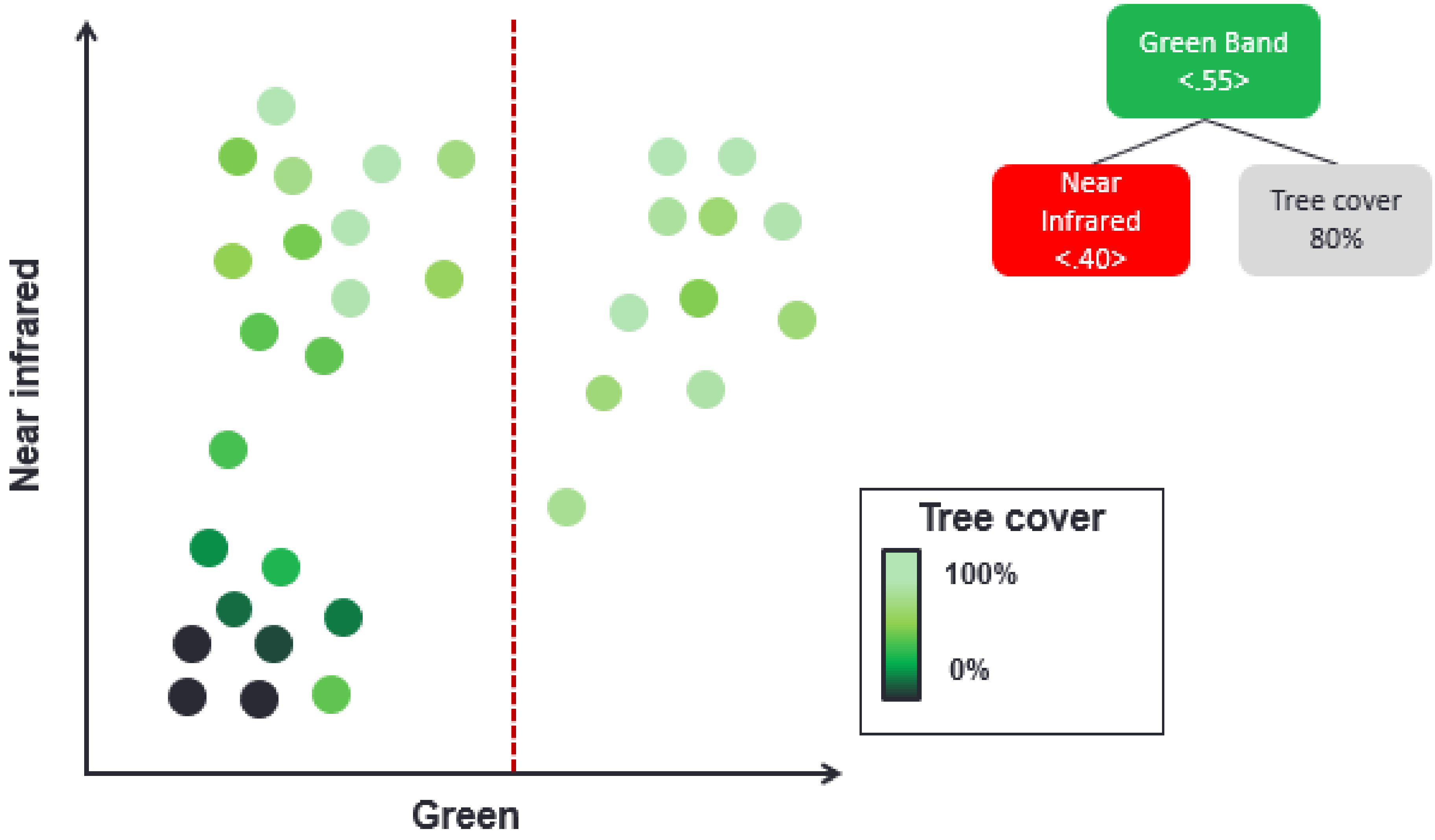
CART Feature Space



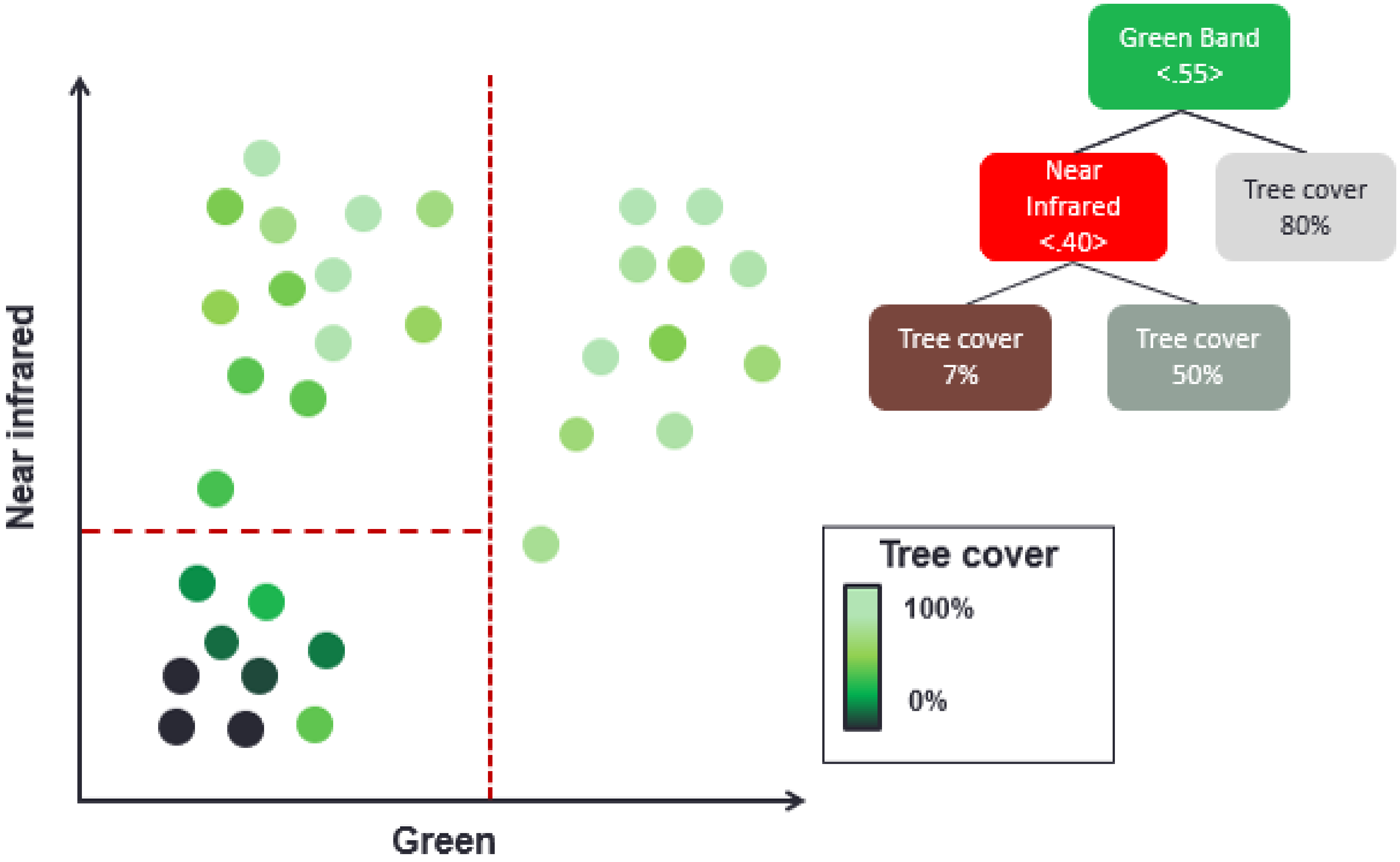
CART Feature Space



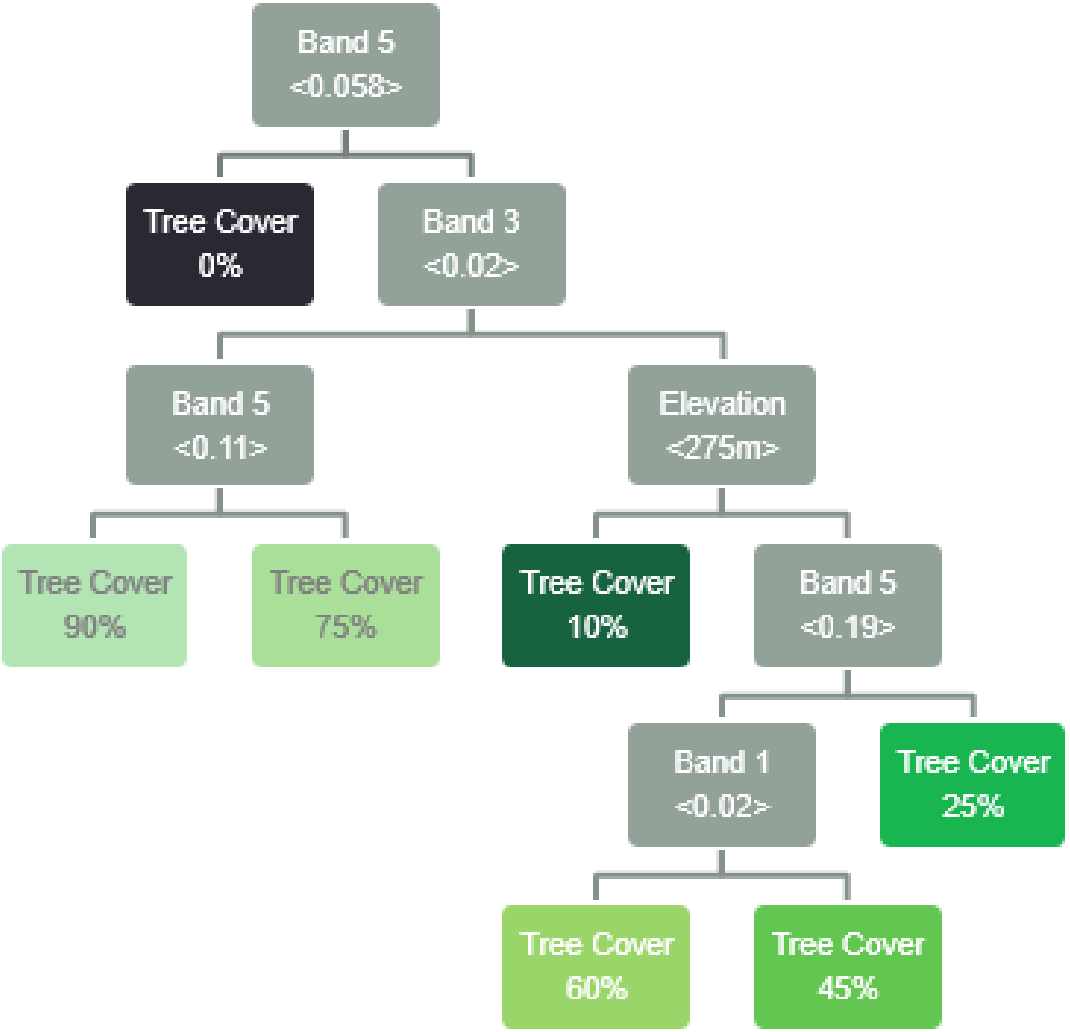
CART Feature Space



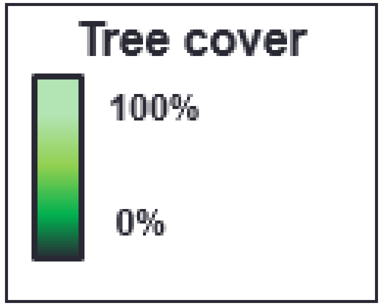
CART Feature Space



Regression tree example



- More informative splits higher up in the tree
- Easily interpretable
- Prunable



Regression vs classification

	Classification trees	Regression trees
Input variables	Categorical	Continuous
Predicted value	Category	mean of the response
Evaluation metrics	Confusion matrix and Kappa	RMSE and R^2

Cons of CART

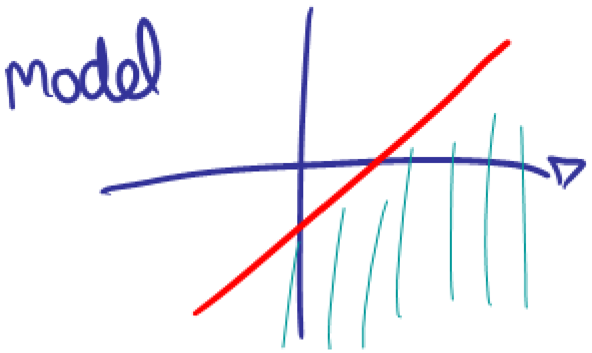
- **Deterministic**
  - Slight changes in data could drastically change model output
- **Bias issue**
  - Some variables have more explanatory power, and they will be chosen over others (which still hold meaningful info)
- **Overfitting**
  - Splits form around the input data
  - Model learns the input data too well
  - Certain decisions may be based on illogical splitting rules (though these can be pruned)

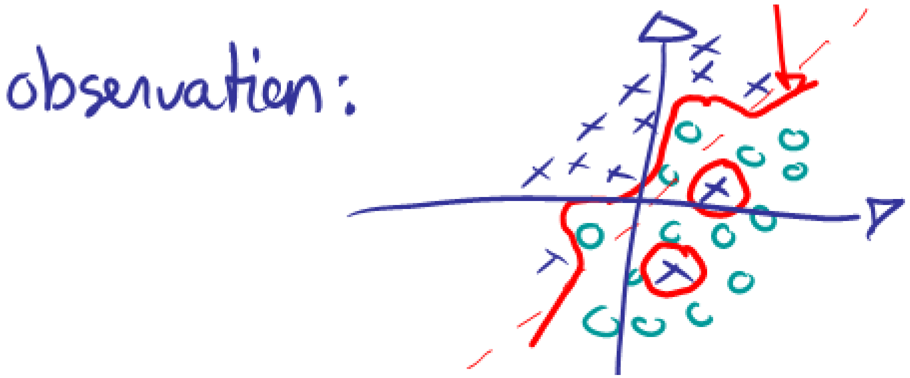
Bias of an estimator

- **Difference between estimated value and actual value**
- **Say I want to predict a certain specific veg type, and I have two variables:**
  - Aspect (limited to N, NE, E, SE, S, SW, W, NW)
  - Near Infrared (DN from 0-255)
- **Which variable is going to give me the most accurate estimate?**

Overfitting

- **Overfitting – doesn't generalize well (or as well as possible)**
  - Given a certain subset, the output model will be biased toward those data
  - Some samples might be more accurate/explanatory than others
- **What does overfitting look like?**



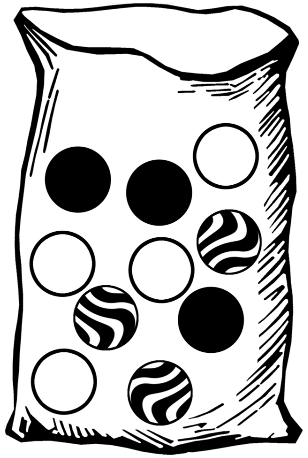


Improving on CART methods

- **Addressing overfitting**
  - May see a pattern in the training data that is not representative of the population
    - some splitting rules may not be informative
    - “correlation does not imply causation”
    - Example:
      - 3 women, 2 men
      - Women are wearing glasses; men are not
      - Use glasses as a splitting rule for gender
  - Incorporating randomization into model helps to minimize the creation of these spurious decision rules

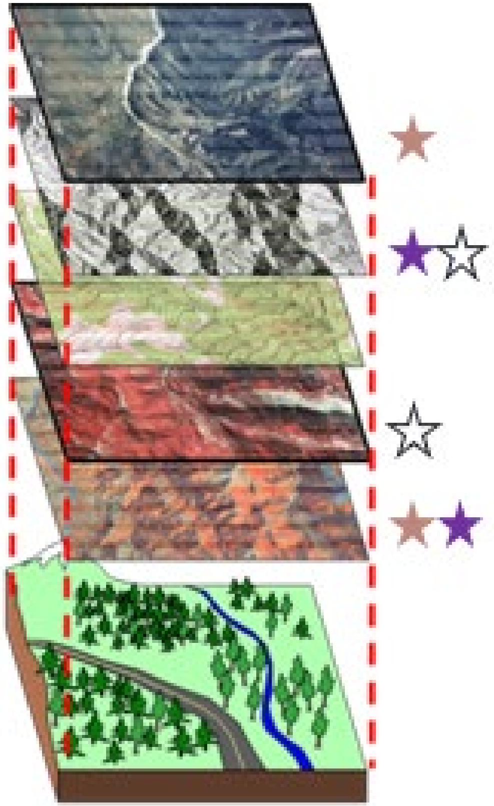
How RF works

- **Bootstrapping**
  - Each tree is created with a unique subsample of the training data, selected with replacement



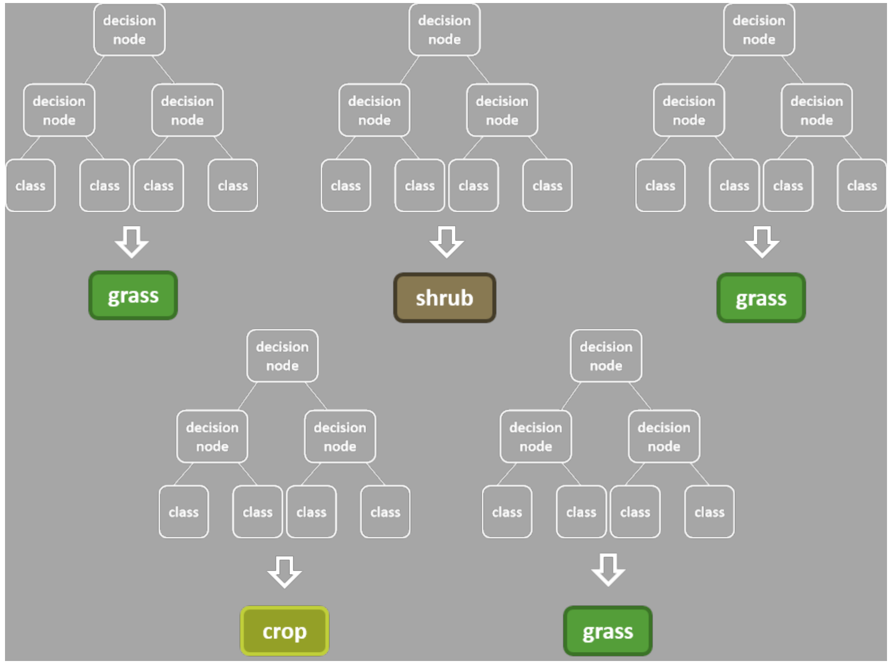
How RF works

- **Bootstrapping**
  - Each tree is created with a unique subsample of the training data, selected with replacement
  - Each decision node uses a random selection of the predictor variables (instead of using all available variables)



How RF works

- **Bootstrapping**
  - Each tree is created with a unique subsample of the training data, selected with replacement
  - Each decision node uses a random selection of the predictor variables (instead of using all available variables)



How RF works

- **Bootstrapping**
  - Each tree is created with a unique subsample of the training data, selected with replacement
  - Each decision node uses a random selection of the predictor variables (instead of using all available variables)
- **Bagging**
  - Each tree = one vote; for each pixel, the majority rules in terms of the output classification

How RF works

- **Bootstrapping**
  - Each tree is created with a unique subsample of the training data, selected with replacement
  - Each decision node uses a random selection of the predictor variables (instead of using all available variables)
- **Bagging**
  - Each tree = one vote; for each pixel, the majority rules in terms of the output classification
- **Each tree in the forest is based on a different subset of data, capturing different phenomena/irregularities**

How RF works – summary

- Lots of decision trees
- Each tree has a unique subset of training data
- Each decision node is based on a random selection of independent variables
- Each tree = 1 vote

Parameterization

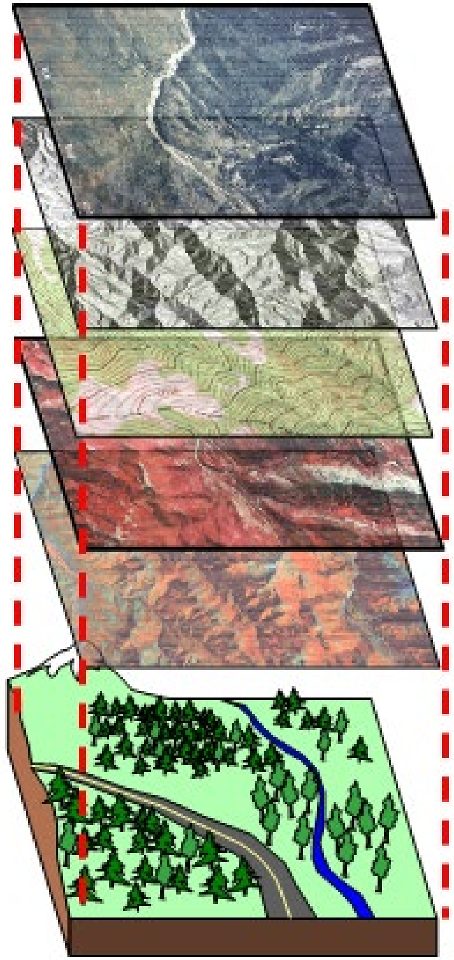
- **# of trees**
  - How many? Eventually reach a saturation point where additional trees do not improve model
- **Variables per split**
  - Usually chosen as the square root of the number of available variables OR set at 2-4
- **Minimum leaf population**
  - The minimum number of pixels classified by a terminal node
- **Bag ratio/fraction**
  - How much of the data should be bagged per tree?

Random Forests as a black box

- **Can't see individual trees / choices**
- **How do we assess which variables are being used? Whether they're being used appropriately?**
- **Variable importance plots**
  - Help us to determine the fit of a model

What goes into models?

- **Training or reference data (point)**
  - Examples of each class (e.g., conifer, aspen, grass, shrub, road, sagebrush, shadow, water, soil, et cetera)
- **Predictor variables**
  - Multispectral imagery
  - Panchromatic imagery
  - Derived variables:
    - NDVI
    - Tasseled Cap transformations (brightness, greenness, and wetness)
  - Topographic variables:
    - Elevation
    - Slope
    - Aspect
  - Bioclimatic variables:
    - Temperature
    - Precipitation
  - Environmental variables:
    - Soils
    - Drainage
    - Land-use
    - Ecoregions



References + further reading about machine learning

- **Random Forests Overview by Breiman and Cutler**
- **Interesting Science Friday segment from 11/20/15**
  - "Why Machines Discriminate—and How to Fix Them" (27:50)
- **Algorithmic Justice League and Coded Bias**

Forest Service
U.S. DEPARTMENT OF AGRICULTURE

Questions?

Random Forests

Juliette Bateman (she/her)
Remote Sensing Specialist/Trainer,
juliette.bateman@usda.gov

Lila Leatherman (they/them)
Remote Sensing Specialist/Trainer
lila.leatherman@usda.gov

Geospatial Technology and Applications Center | GTAC
USDA Forest Service