Random Forest: A Short Introduction

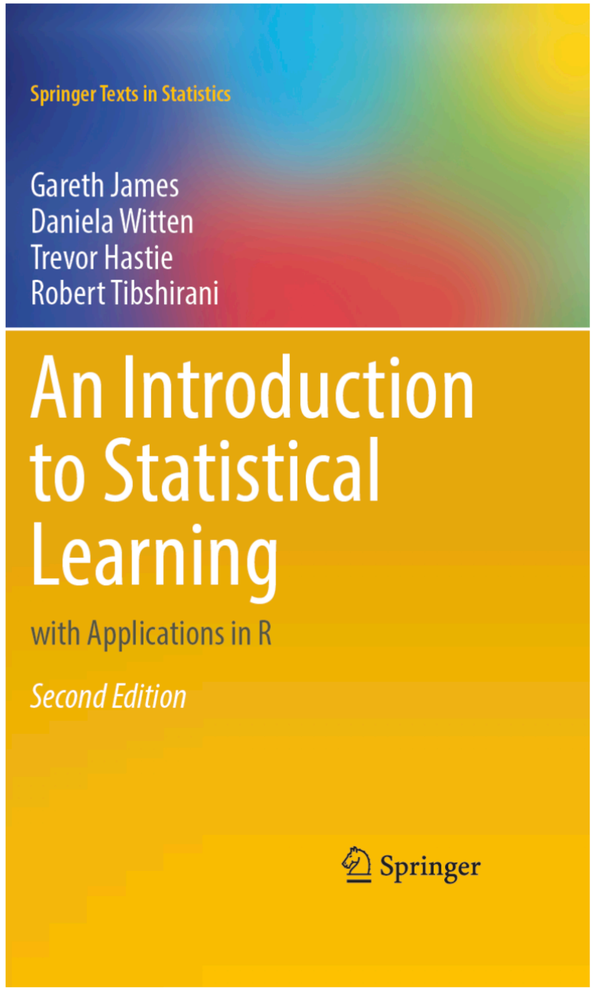
Bilgecan Şen

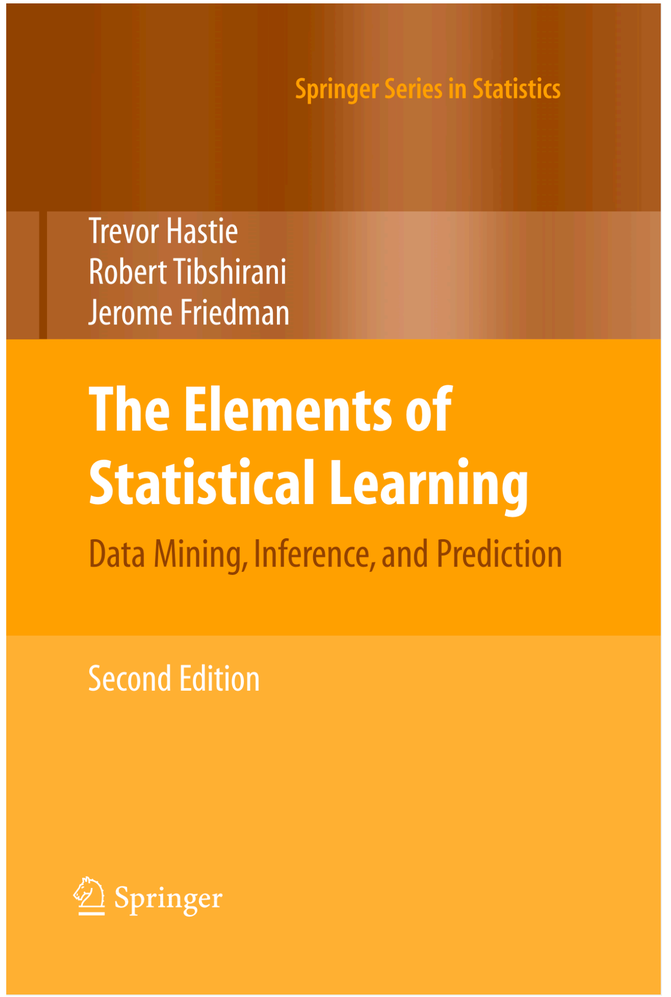
University of Maryland
CENTER FOR ENVIRONMENTAL SCIENCE
APPALACHIAN LABORATORY

A bit about me

- Assistant research scientist at UMCES Appalachian lab
- Quantitative ecologist
- Predictability, forecasting, macroecology
- https://github.com/bilgecansen

Suggested textbooks





Why random forests?

- Higher predictive accuracy compared to linear regression
- More resistant to overfitting (p >> n)
- Can accommodate non-linear relationships
- Non-parametric, no need to know beforehand the underlying relationships
- Cross-validation baked in to model fitting

Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Abundace ~ Temp + Prec

Abundance	Temperature	Precipitation
33	1.75	208.08
52	5.92	162.78
50	7.81	190.72
27	-1.45	191.66
69	6.77	232.81
30	3.05	129.42
51	9.53	182.53
69	10.61	231.58
55	8.75	179.33
48	6.81	155.54

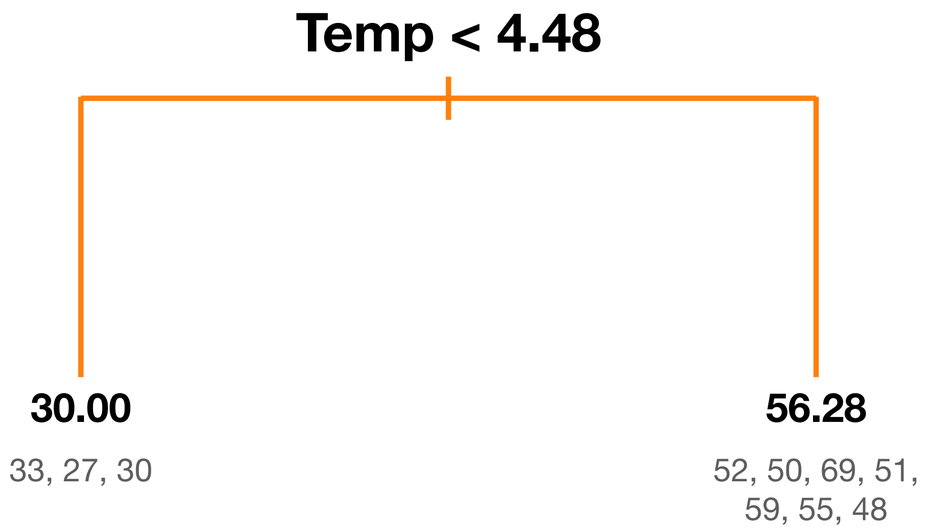
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Abundance ~ Temp + Prec



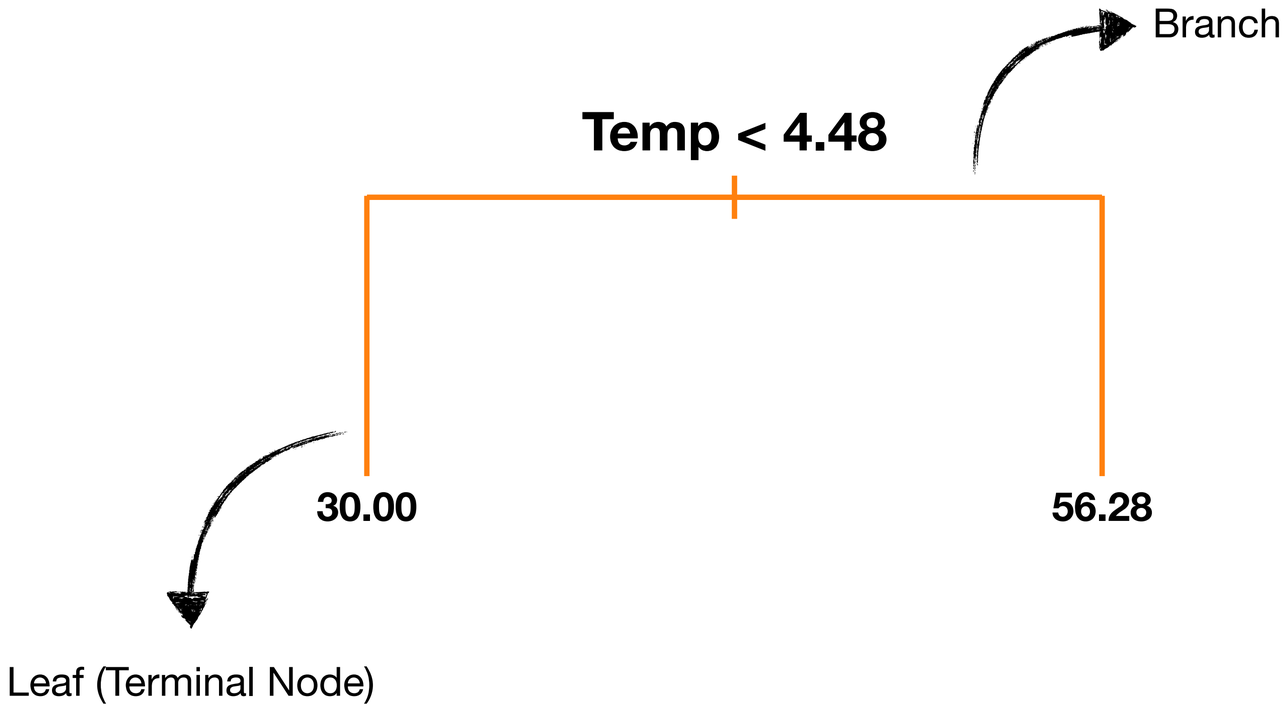
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Abundance ~ Temp + Prec



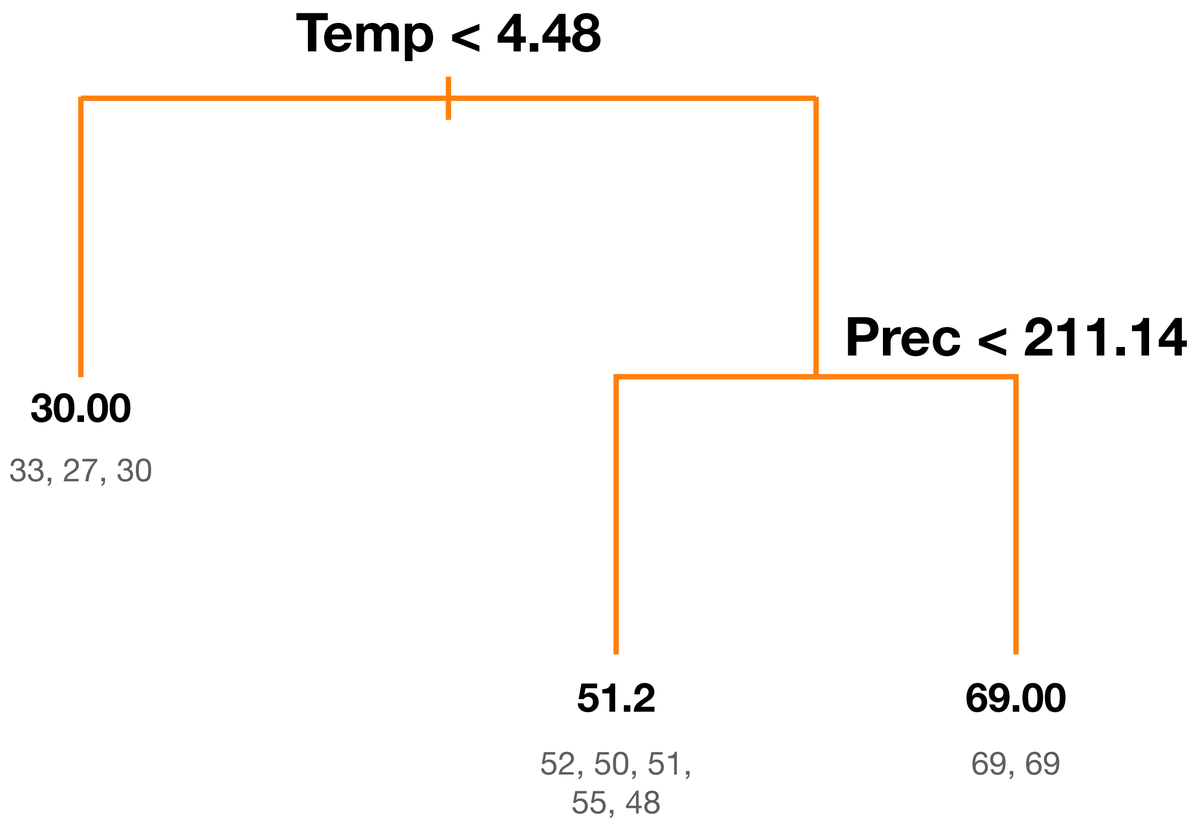
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Abundance ~ Temp + Prec



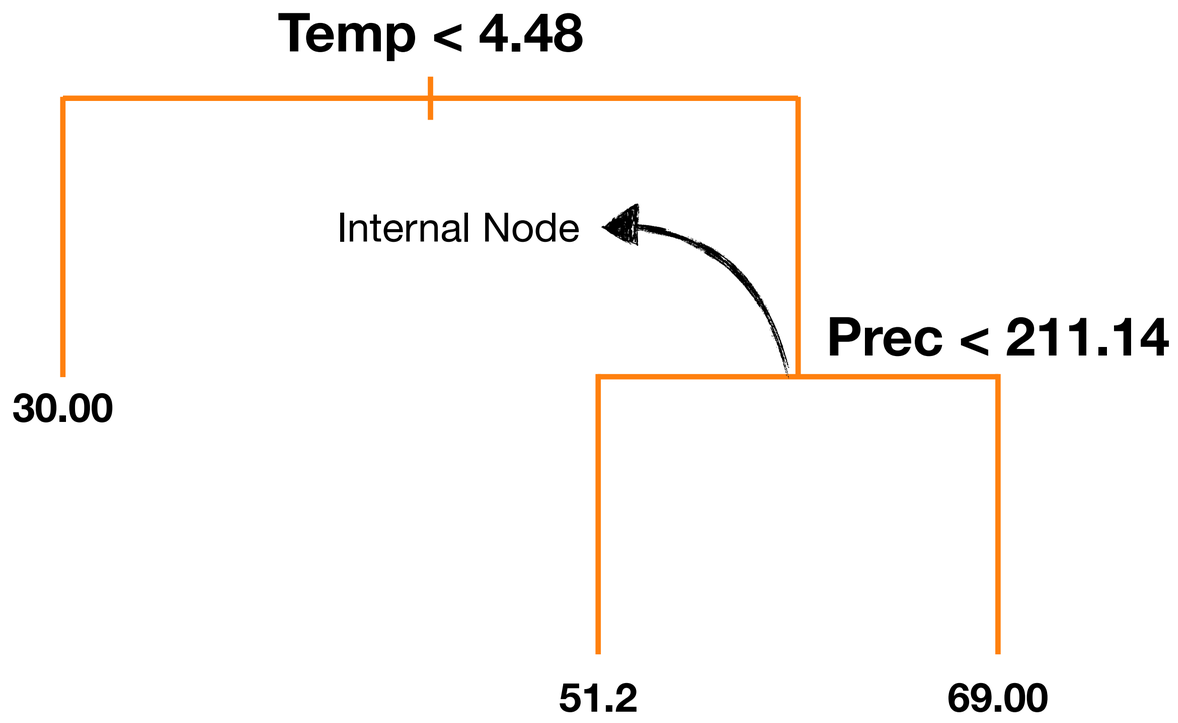
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

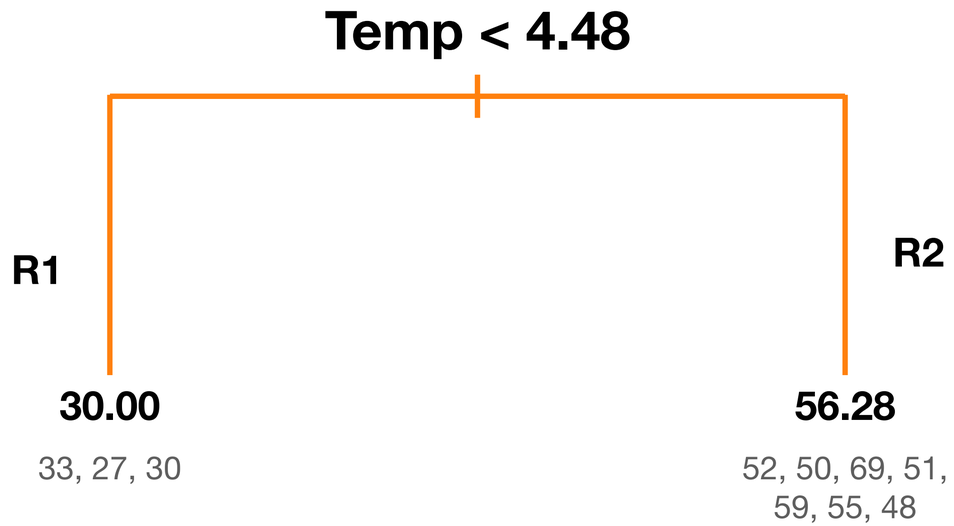
- **Simple example:**

  Abundance ~ Temp + Prec



Classification and Regression Trees (CART)

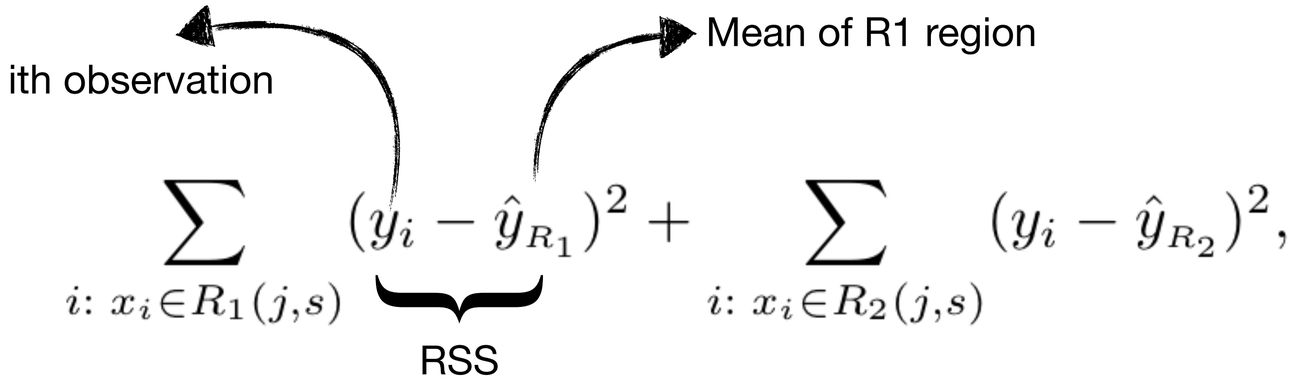
- How do we decide where to split the data?



$$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\},$$

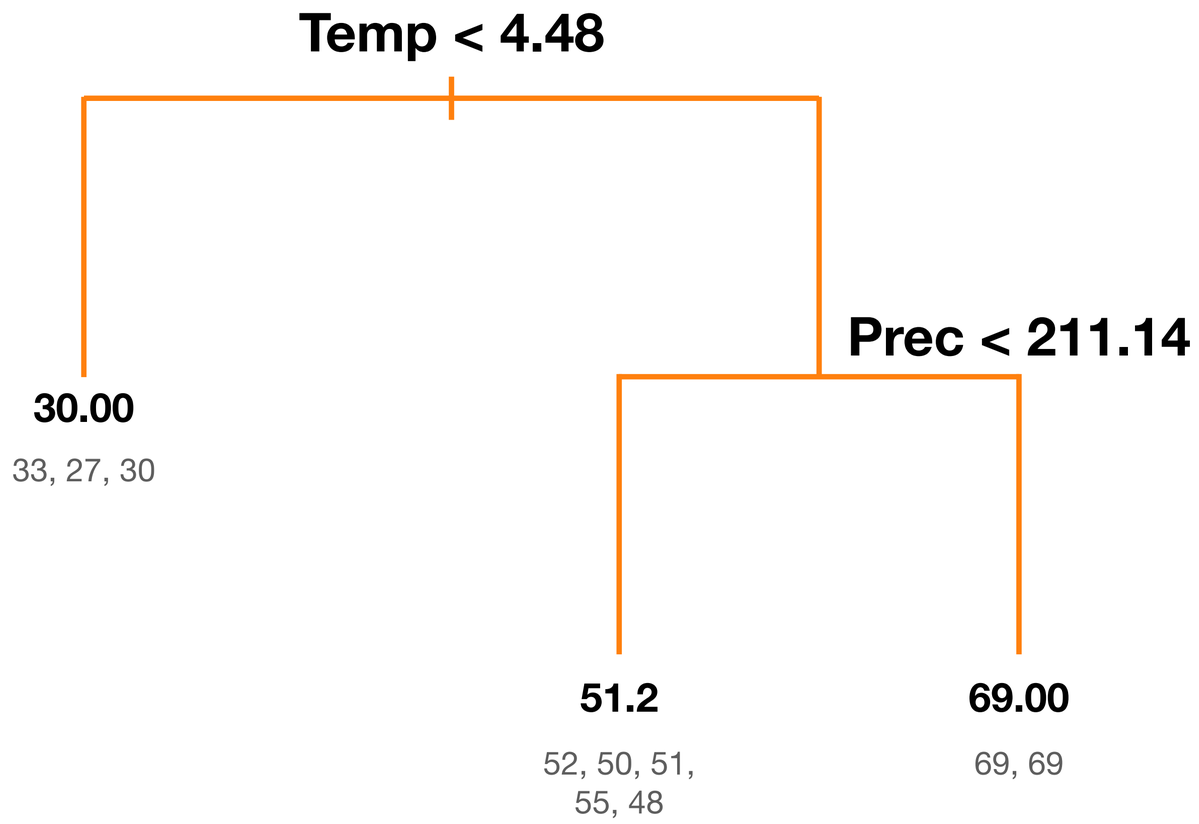
Classification and Regression Trees (CART)

- How do we decide where to split the data?
- Find the split that minimizes total residual sum of squares in split regions



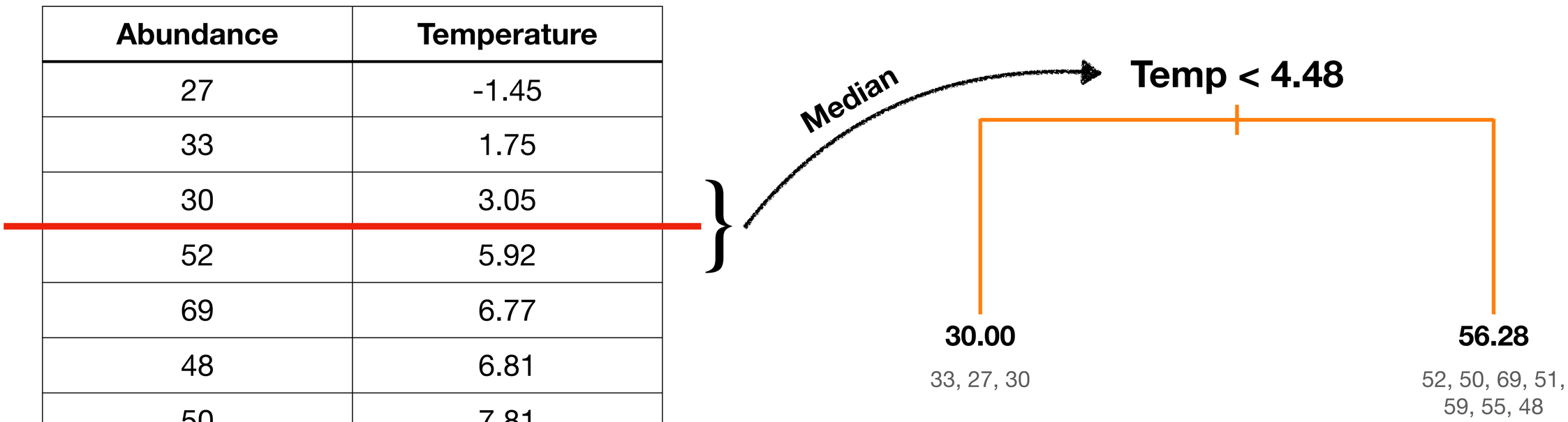
Classification and Regression Trees (CART)

- How do we decide which variables to use in each split?
- This is a greedy algorithm
- In all splits, try all variables and all their potential splits



Classification and Regression Trees (CART)

Abundance	Temperature
27	-1.45
33	1.75
30	3.05
52	5.92
69	6.77
48	6.81
50	7.81
55	8.75
51	9.53
69	10.61

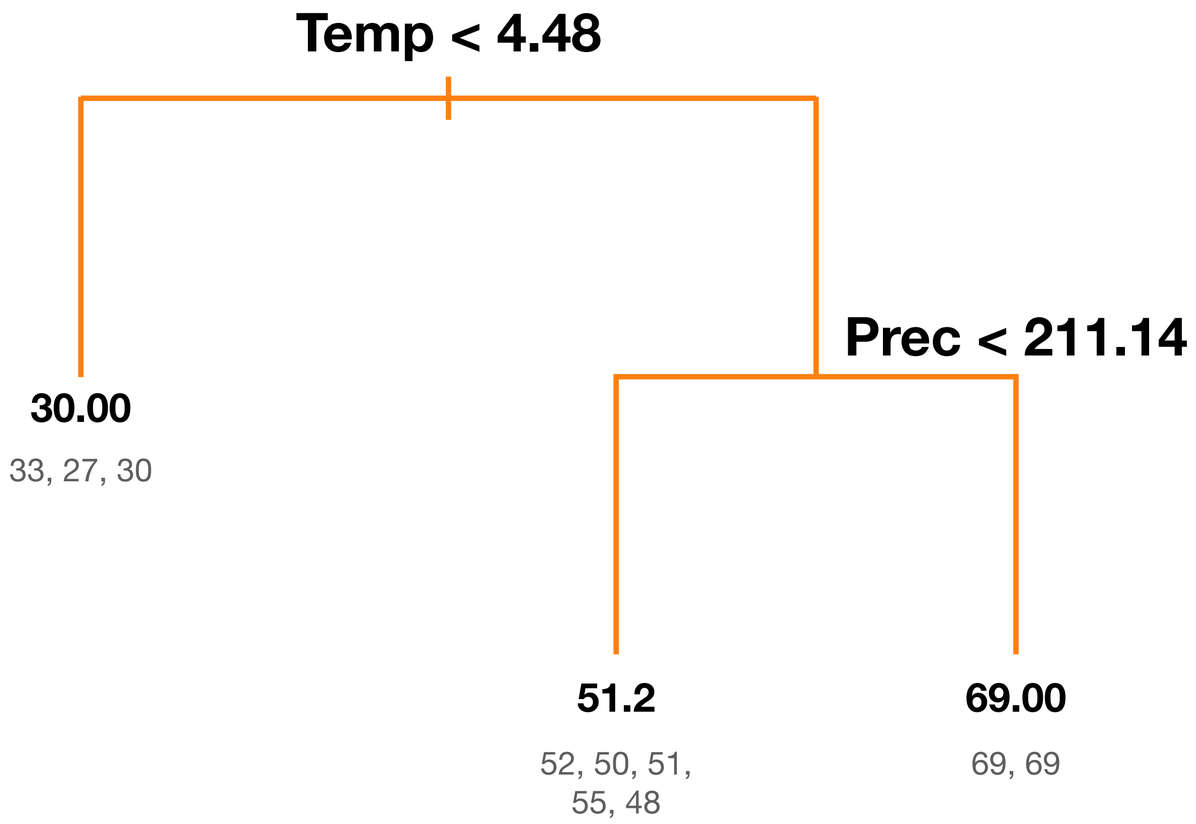


Classification and Regression Trees (CART)

- When to stop growing a tree?

  1. A predetermined rule

  2. Grow as large as possible, then prune



Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Occurence ~ Temp + Prec

Occurence	Temperature	Precipitation
0	1.75	208.08
1	5.92	162.78
1	7.81	190.72
0	-1.45	191.66
1	6.77	232.81
0	3.05	129.42
1	9.53	182.53
1	10.61	231.58
0	8.75	179.33
1	6.81	155.54

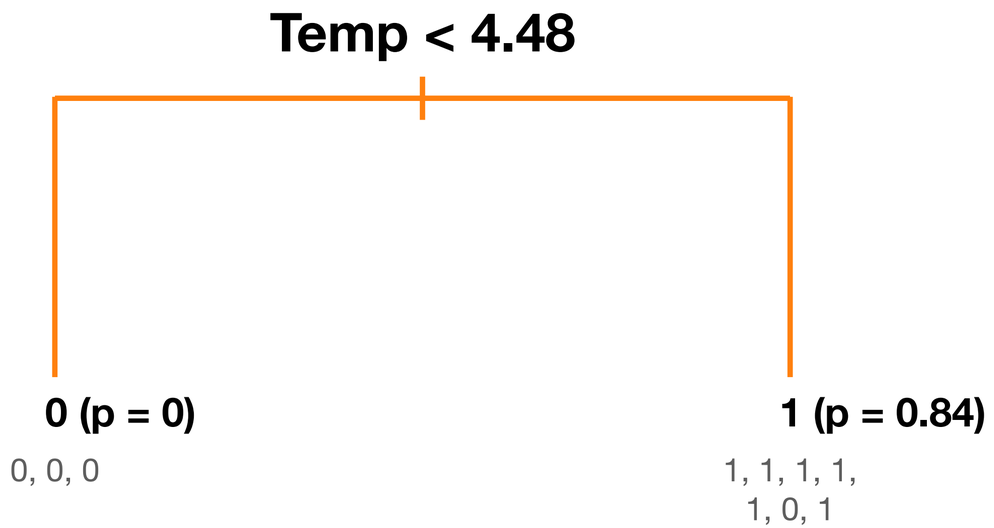
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

  Occurence ~ Temp + Prec



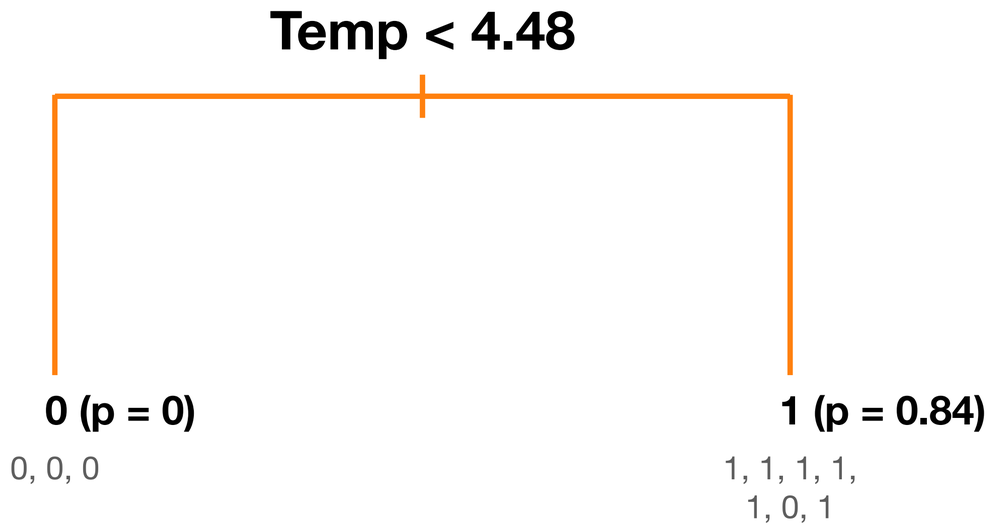
Classification and Regression Trees (CART)

- **Core method:**

  Recursive binary splits

- **Simple example:**

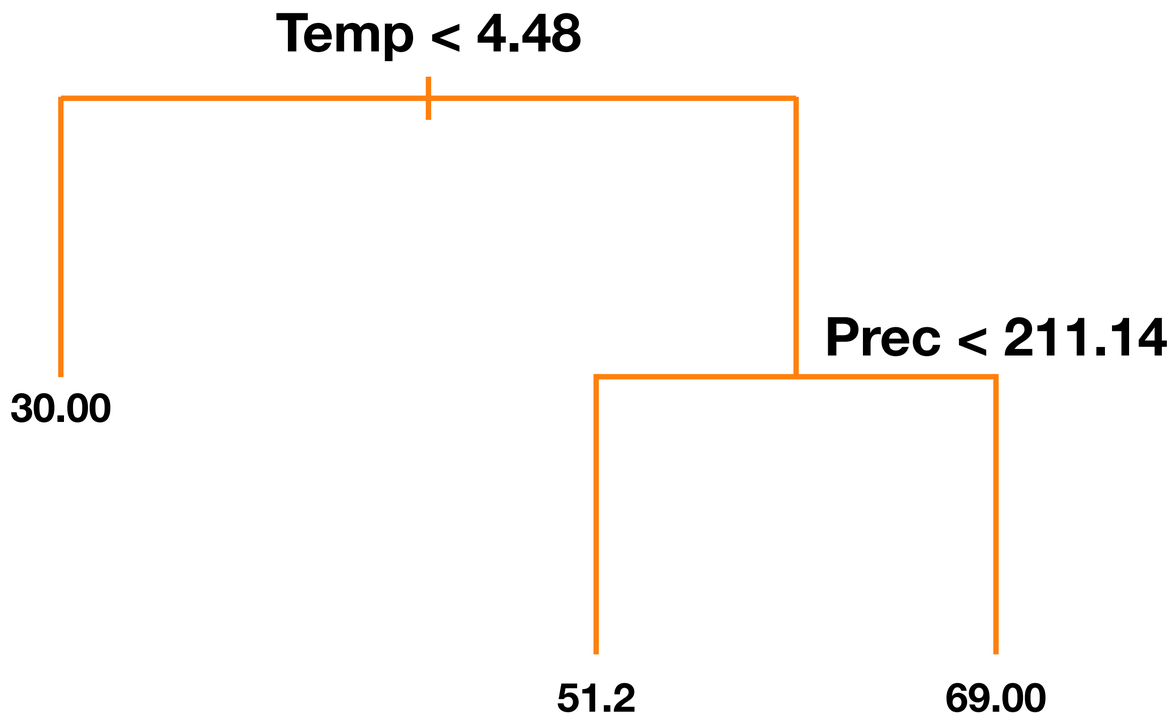
  Occurence ~ Temp + Prec



$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

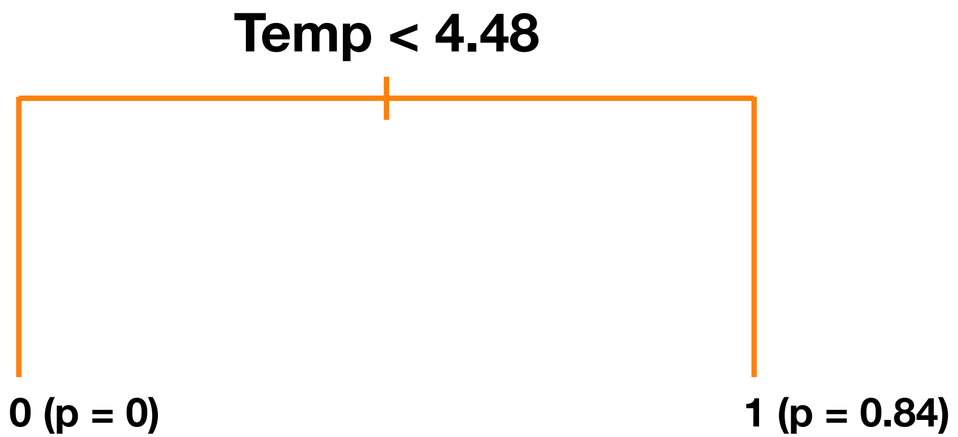
Predictions with CART

- Temp = 5, Prec = 300
- Temp = 3, Prec = 250
- Temp = 6, Prec = 100



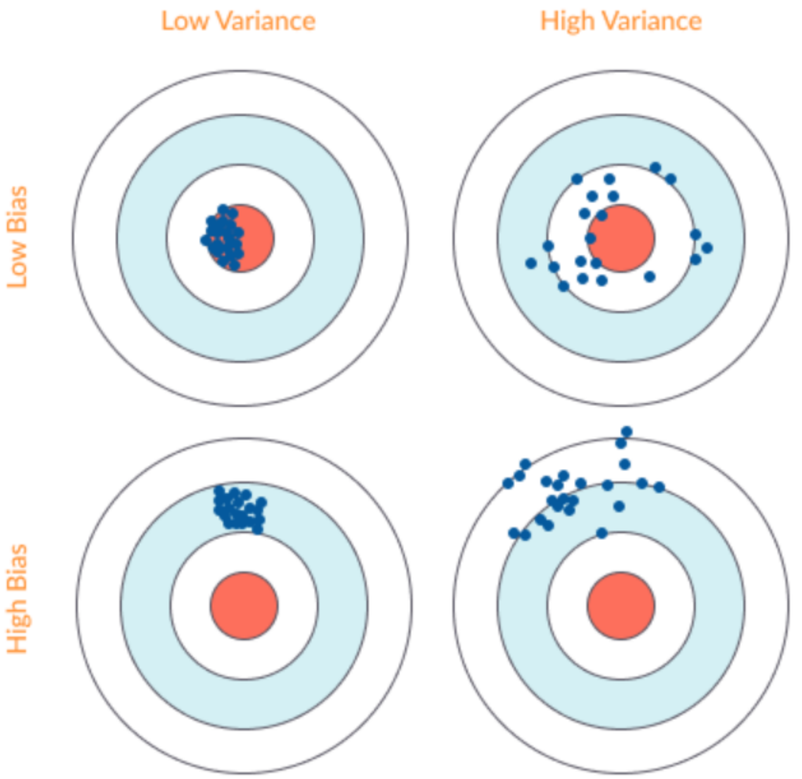
Predictions with CART

- Temp = 5
- Temp = 3

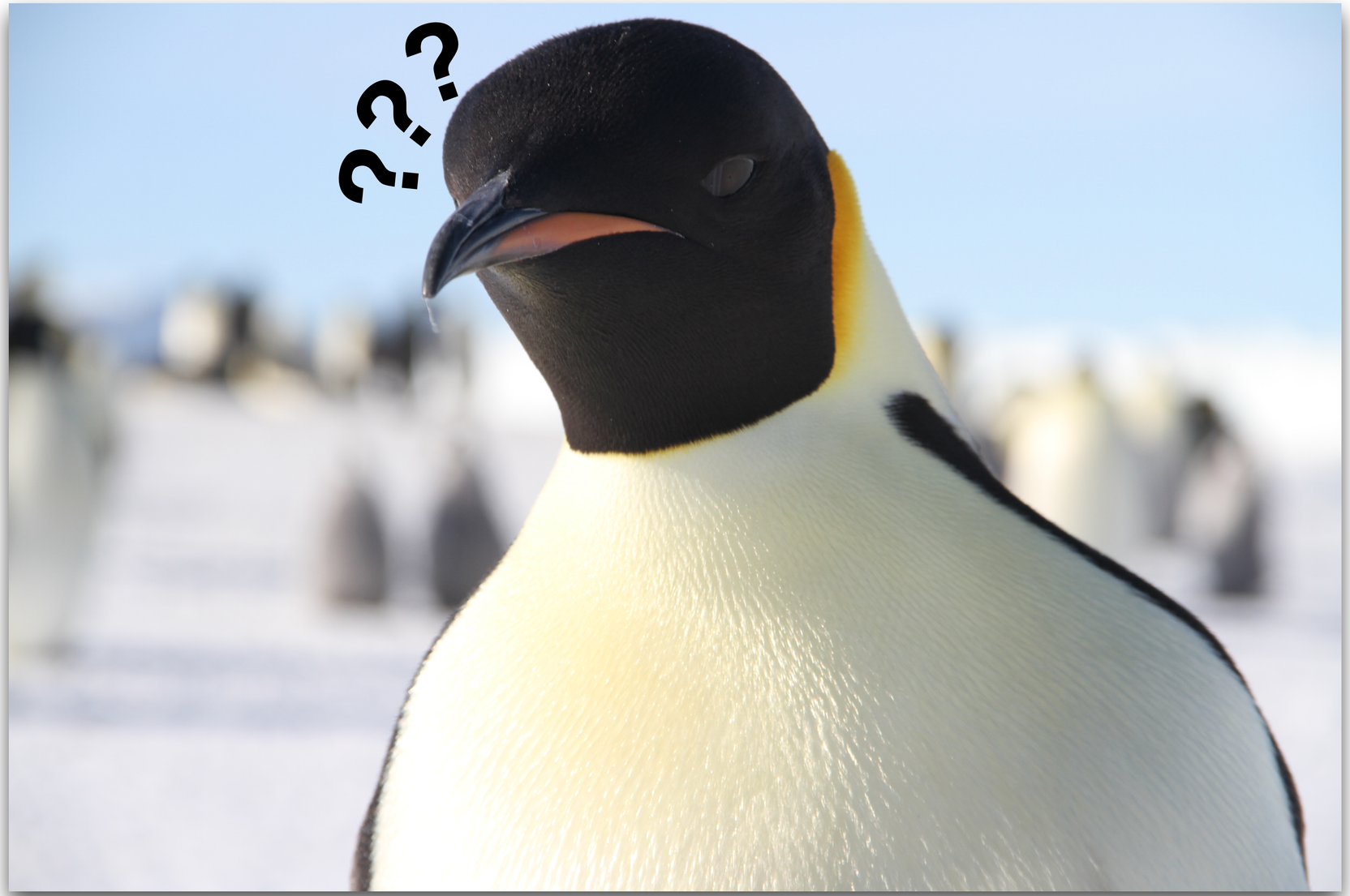


Issues with CART

- High variance
- Low prediction accuracy
- How to overcome high variance?



Questions?

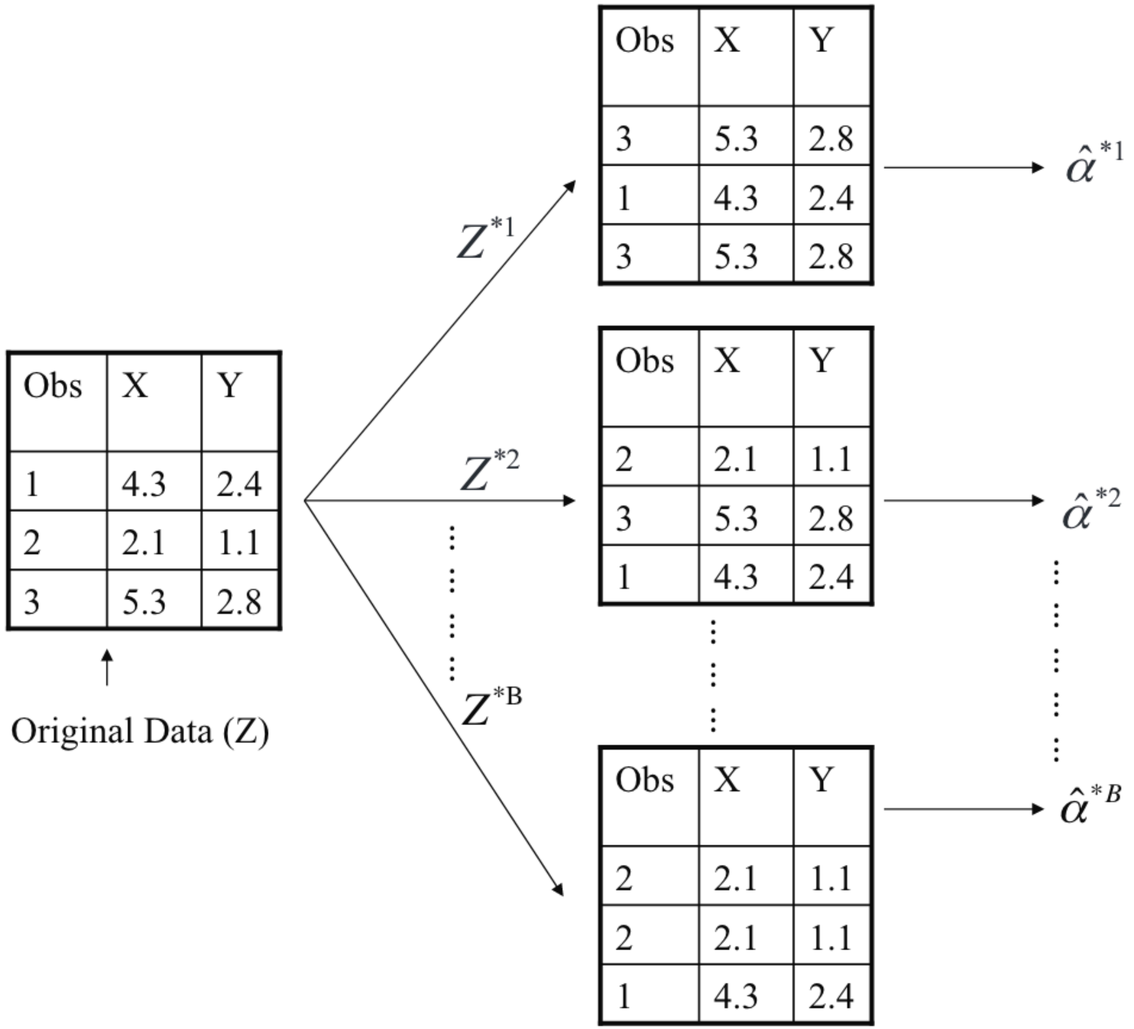


Bagging

- Bootstrap aggregation
- Bootstrap data many times, then fit a CART to each one

Bagging - Bootstrap

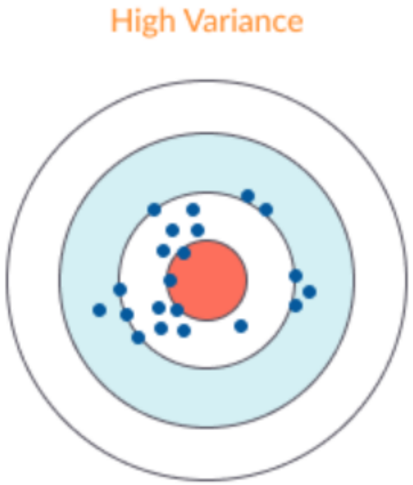
- Sampling with replacement



Bagging

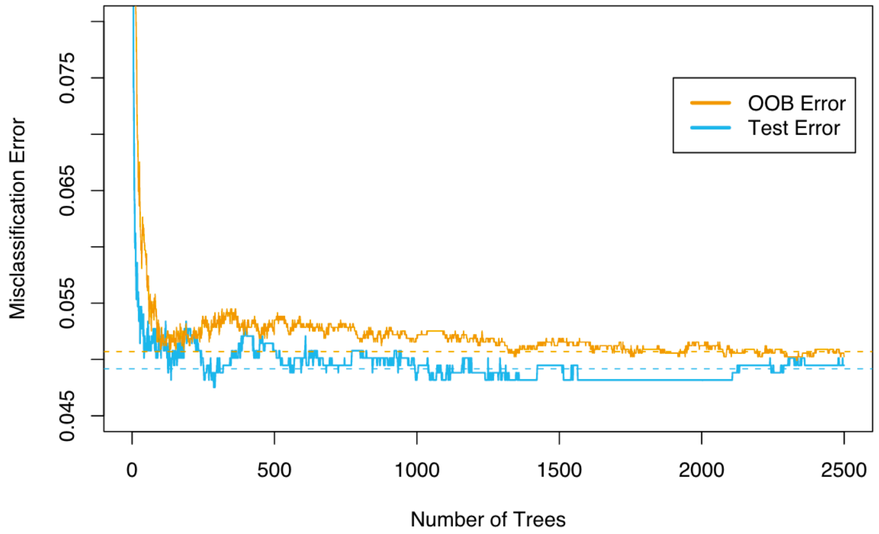
- Bootstrap aggregation
- Bootstrap data many times, then fit a CART to each one
- Average across CARTs to obtain a single low-variance model

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



Bagging - OOB

- On average bootstrapping only result in 66% of the original data in each set
- Unused data can be used as a test set
- This is equivalent to cross validation
- Out-of-bag RMSE, R2 (regression) or error rate (classification)



Bagging - OOB

- Predict each observation using trees it was absent (approximately 1/3 of trees)
- Then,

  Regression: average across these trees to get a single OOB prediction

  Classification: Majority votes across to get a single OOB prediction

Bagging may not be enough

- Each model fit to a bootstrapped sample should be independent
- Trees are correlated!
- All trees might end up looking very similar (e.g., strongest predictor will always be at the top of the tree)

Random forest

- Random forest uses bagging but decorrelates trees
  - In each split a random set of m predictors are chosen as split candidates
  - Moderately strong predictors can now appear at the top of the tree

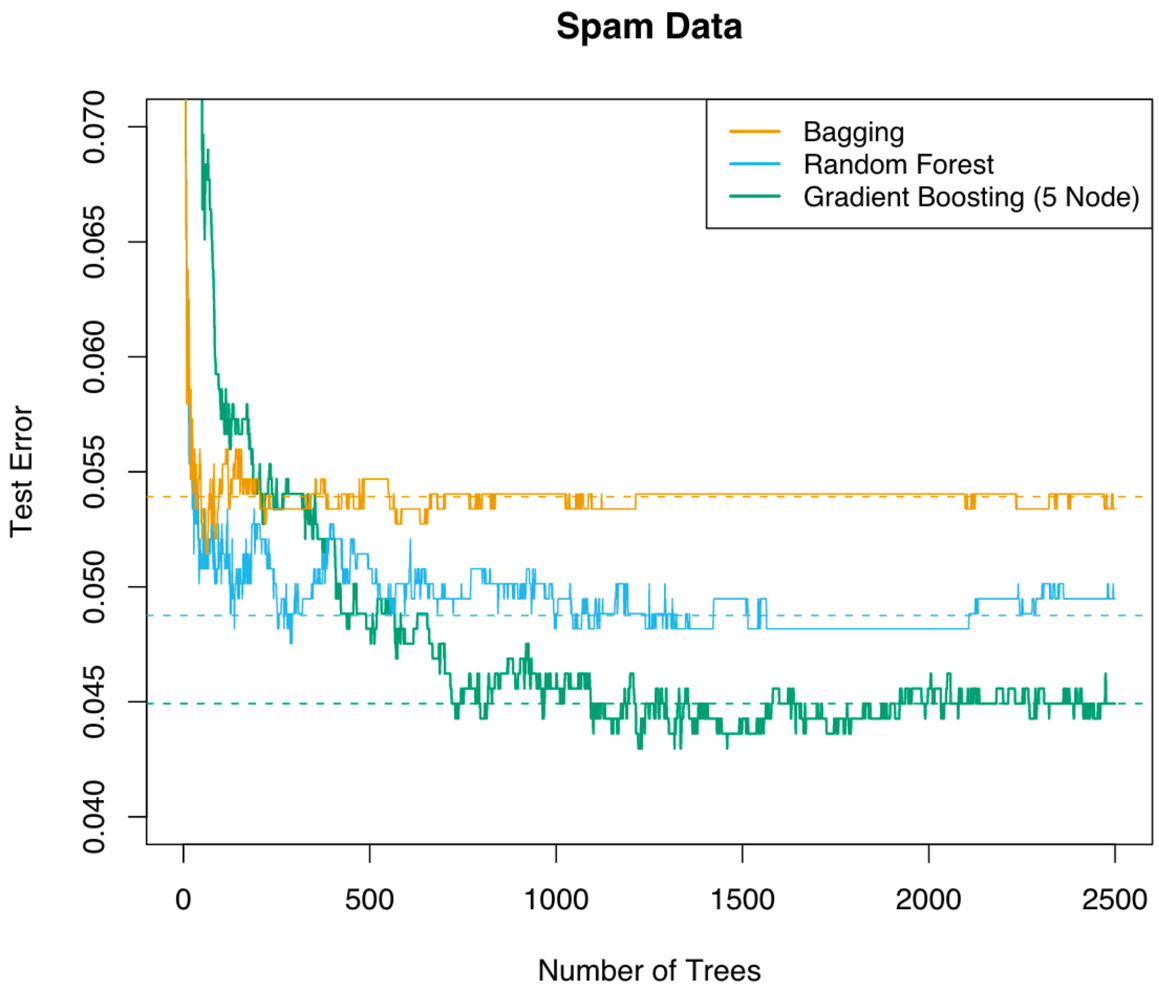
Random forest

- Two parameters to tune for RF
  - m, number of randomly picked candidate variables
  - Number of trees
- Default m is $\sqrt{p}$ for regression, $p/3$ for classification
- Number of trees should be as reasonably high as possible (usually more than 1000)

Random forest - Steps

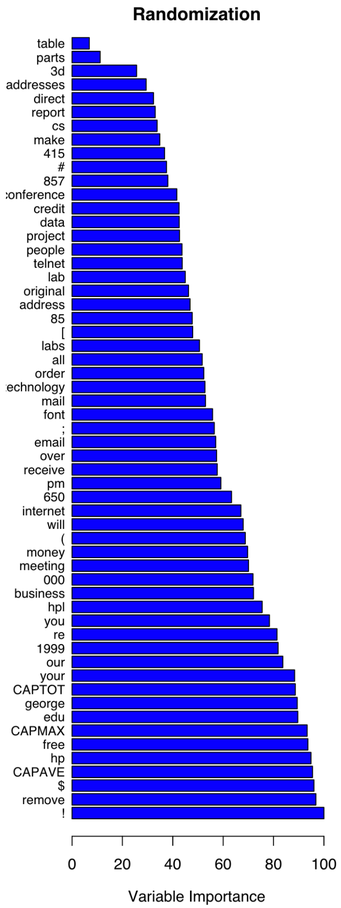
1. Bootstrap the training data many times
2. Fit a CART to each bootstrapped sample
   - In each split contest randomly determined m variables
   - Grow the tree as large as possible (no pruning)
3. Predict OOB data
   - **Regression:** Average across trees
   - **Classification:** Majority vote across trees

Random forest - Performance

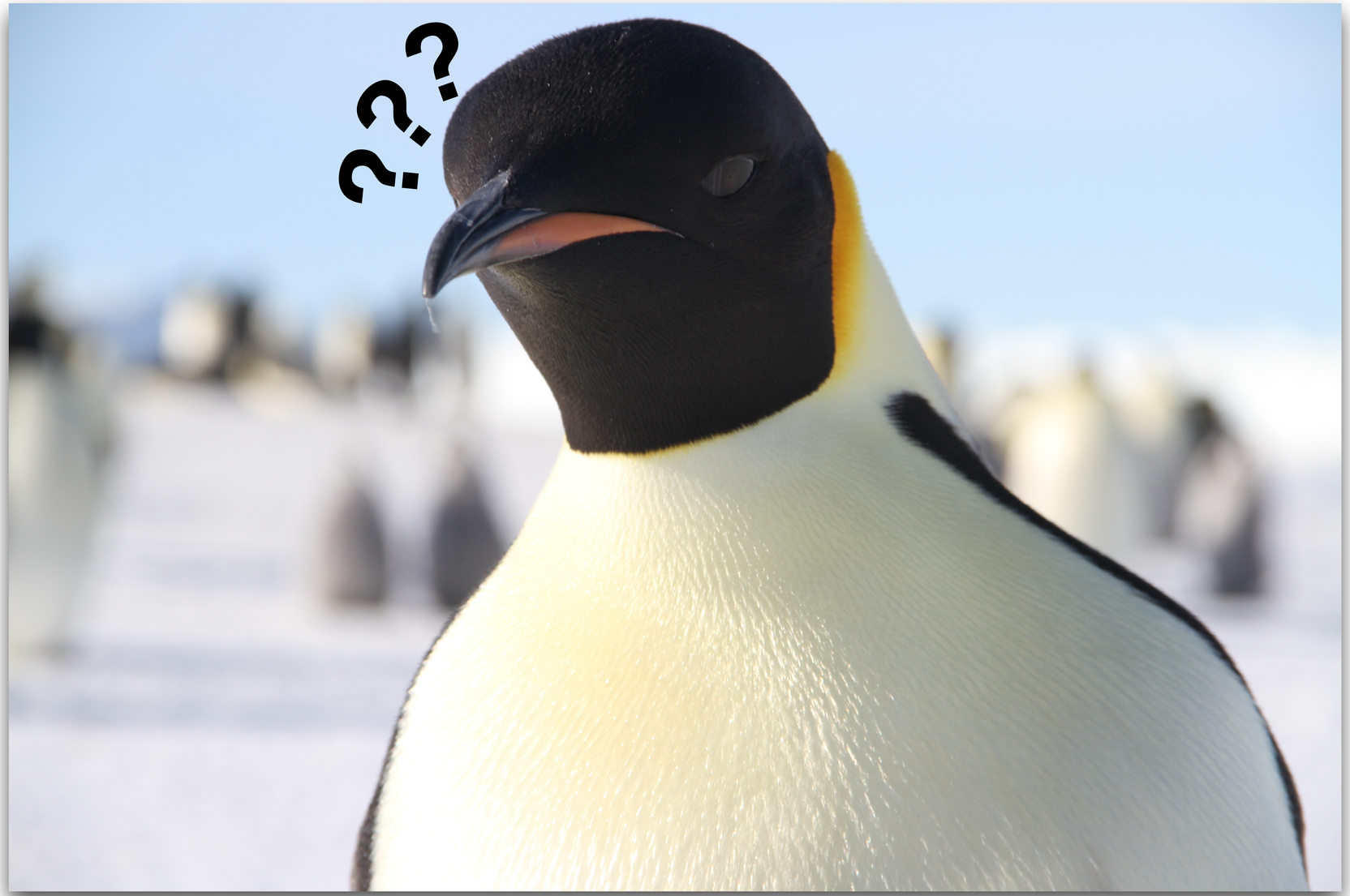


Random forest - Variable Importance

- Permute (randomize) each variable
- Measure the decrease in OOB-RSS (regression) or OOB-error rate (classification)

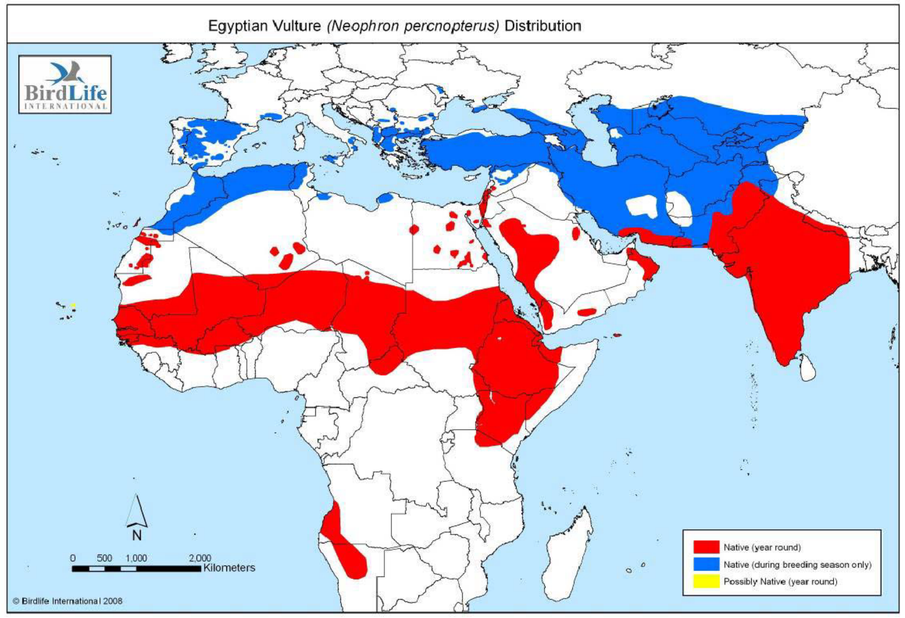


Questions?



Lab: Nesting habitat of Egyptian vulture

- Egyptian Vulture (*Neophron percnopterus*)
  - IUCN red list category: Endangered

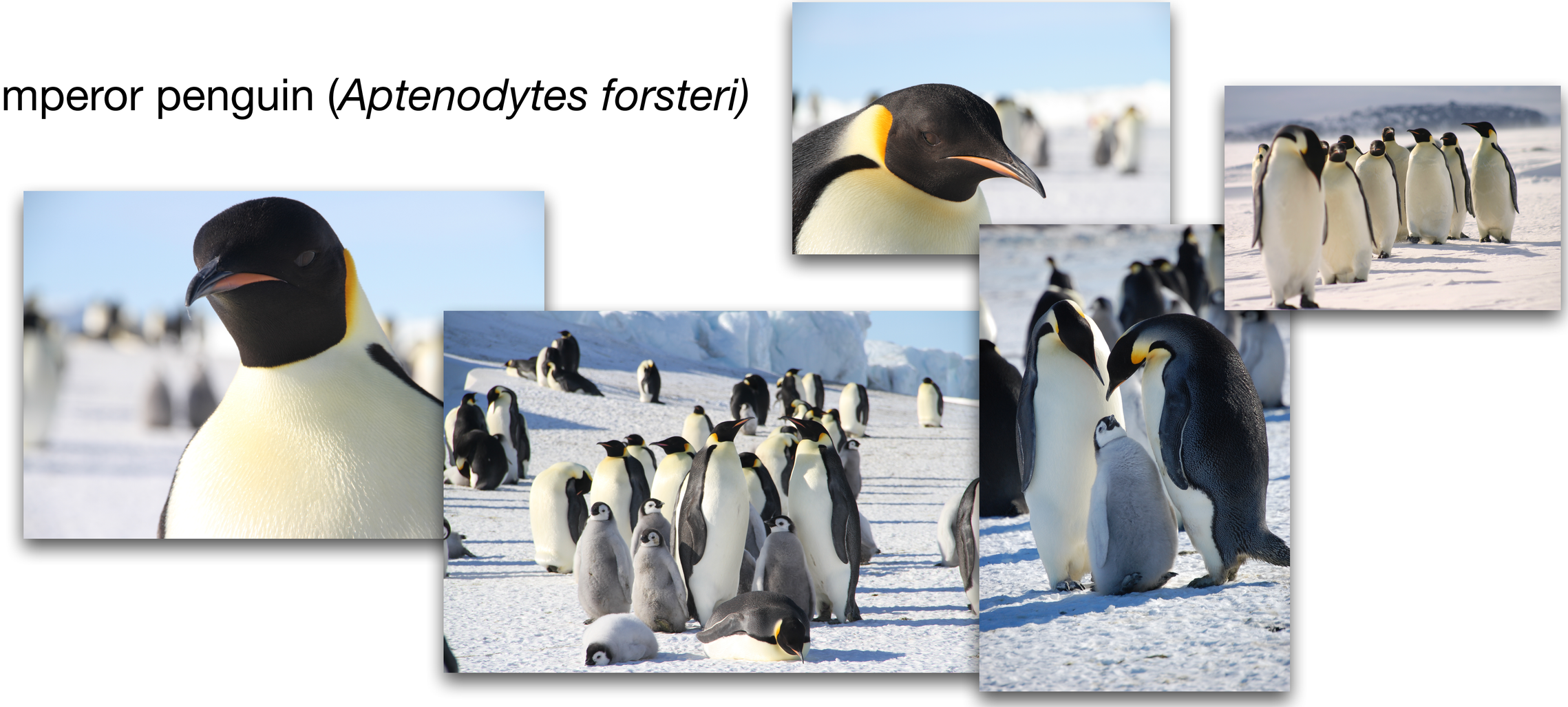


Lab: Nesting habitat of Egyptian vulture

- Nest occurrence in Beypazari, Turkey
- 39 nests, 30 absence points
- 21 Variables
  - Elevation
  - Distance to nearest village
  - Percent forest in 1 km radius

Lab: Colony abundance of emperor penguin

- Emperor penguin (*Aptenodytes forsteri)*



Lab: Colony abundance of emperor penguin

- Emperor penguin (*Aptenodytes forsteri)*
  - Average colony abundance between 2009-2018
  - 32 variables
    - Wind speed
    - Sea-ice concentration
    - Mixed layer depth
    - Distance to nearest fastice edge