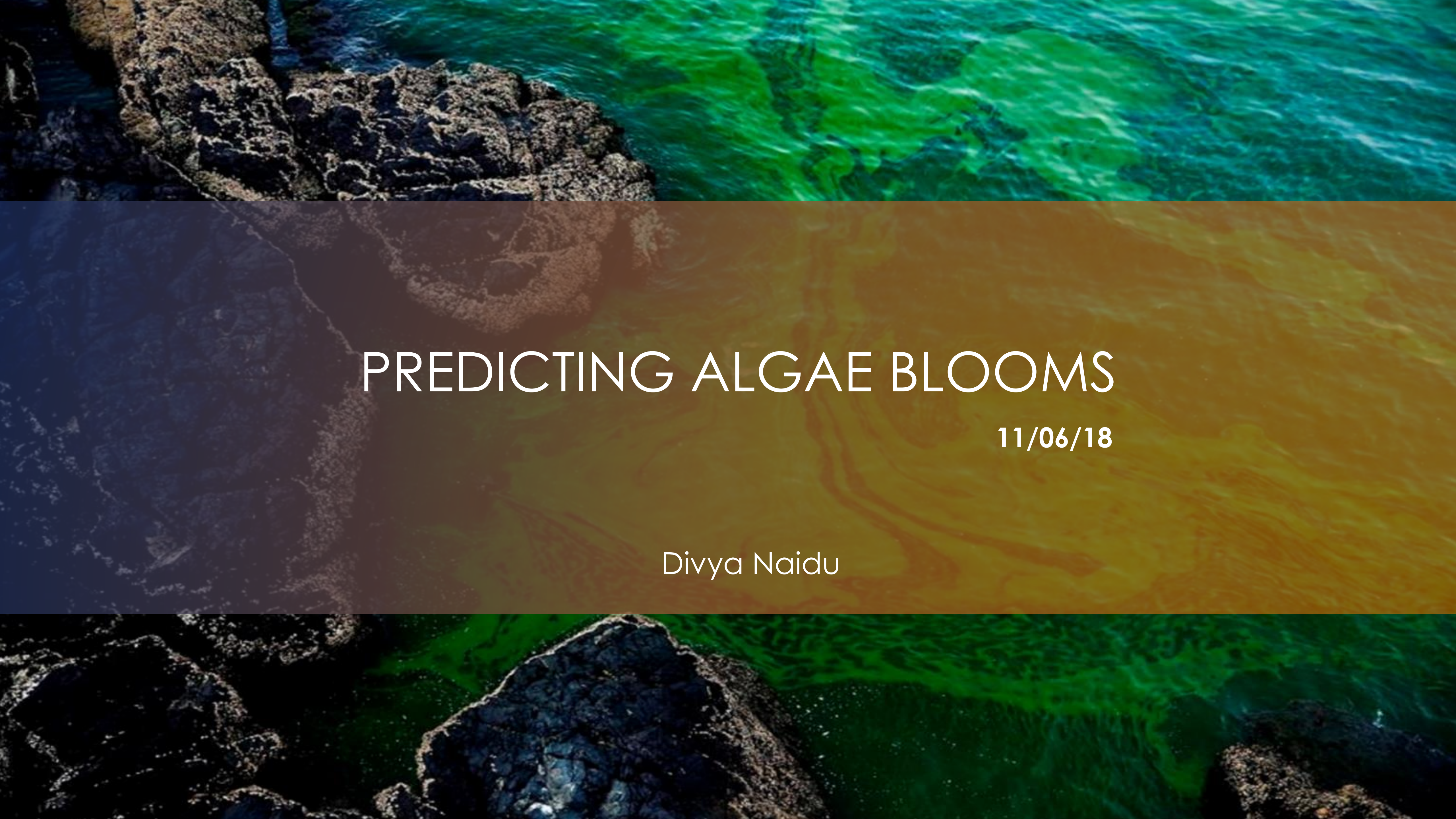
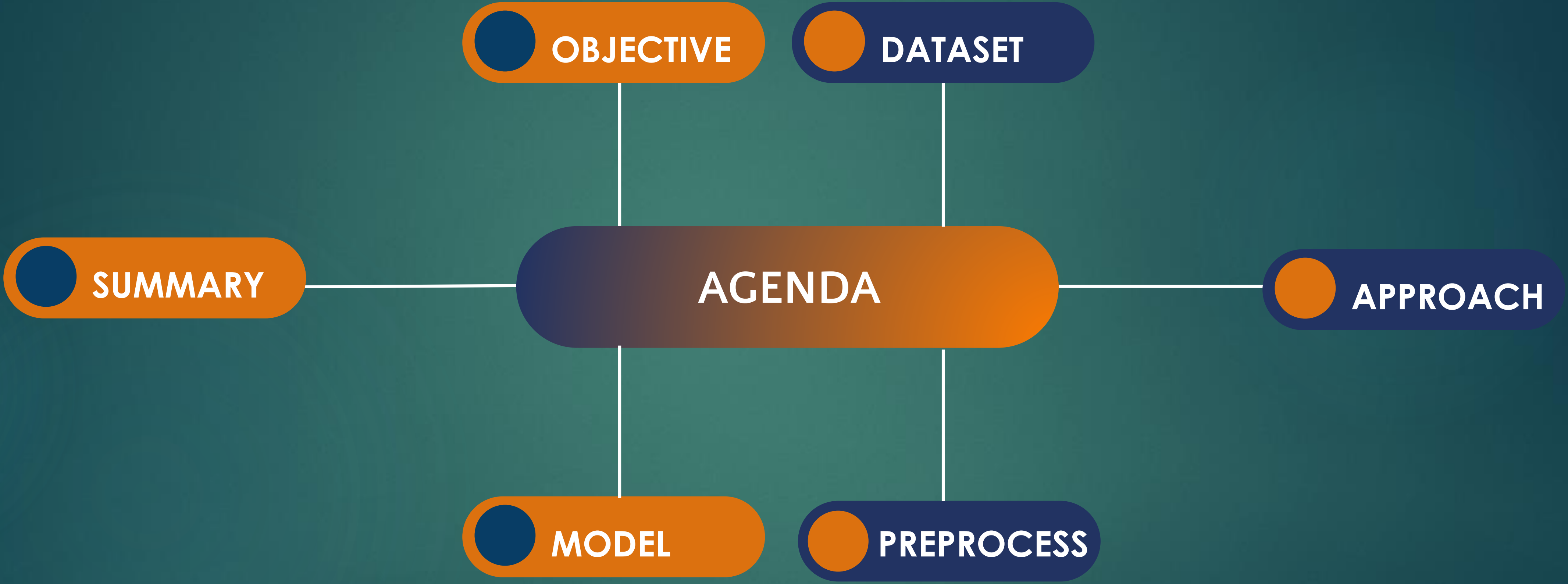


PREDICTING ALGAE BLOOMS

11/06/18

Divya Naidu

OUTLINE



OBJECTIVE

DATASET

SUMMARY

AGENDA

APPROACH

MODEL

PREPROCESS

OBJECTIVE

SIGNIFICANCE

High Concentrations of harmful algae in rivers constitute serious ecological problem

METHODOLOGY

Biological Analysis - Expensive

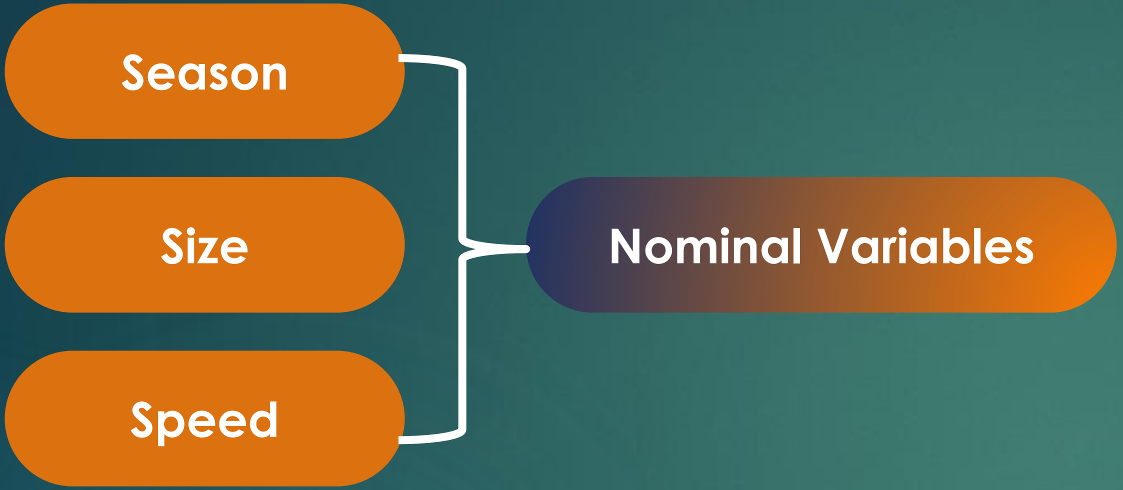
Chemical Monitoring - Cheap

GOAL

How environmental factors influence algae frequencies?

Build a Model to Monitor and Perform an early forecast of algae blooms to improve quality of river

DATA SETS



Season

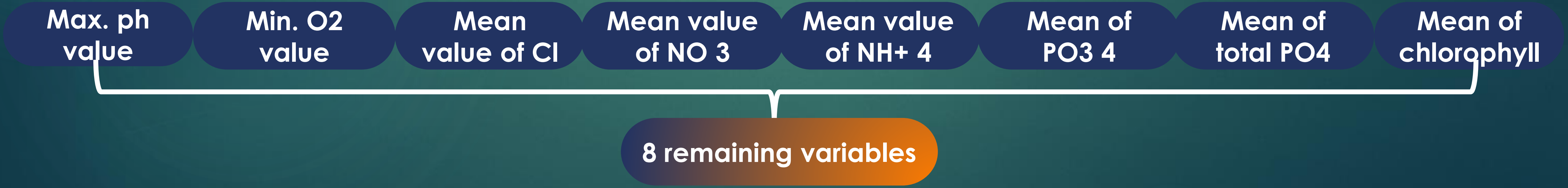
Size

Speed

Nominal Variables

The first dataset is train data - contains information on 200 water samples with 11 variables.

The second dataset is test data - contains information on 140 extra observations.



Max. ph value

Min. O2 value

Mean value of Cl

Mean value of NO 3

Mean value of NH+ 4

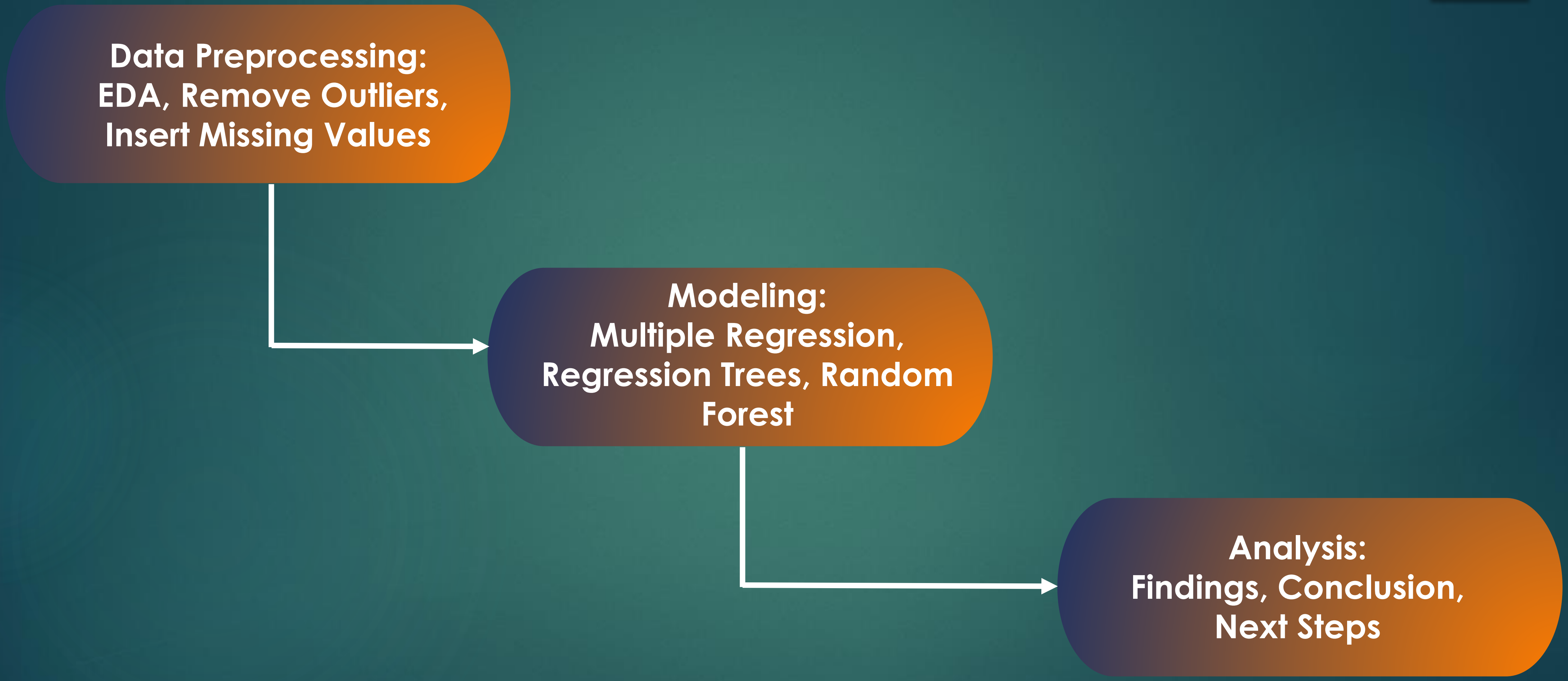
Mean of PO3 4

Mean of total PO4

Mean of chlorophyll

8 remaining variables

APPROACH

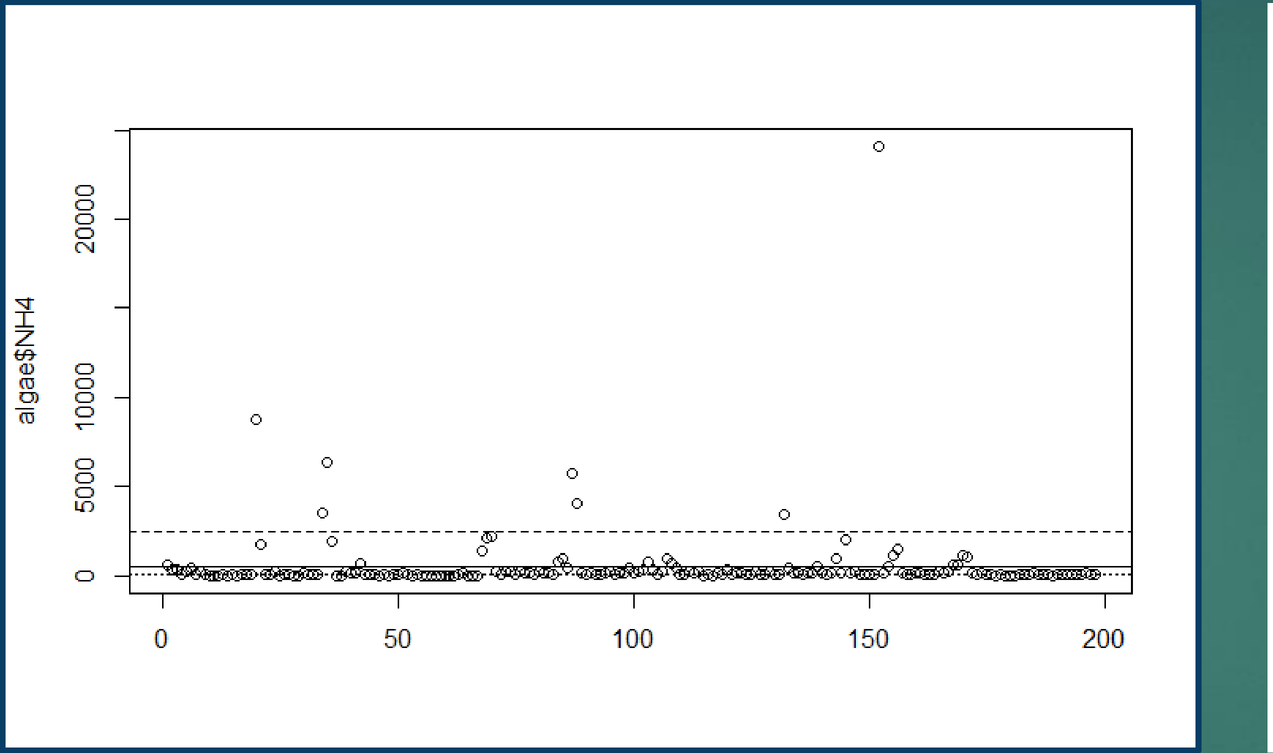


Data Preprocessing:
EDA, Remove Outliers,
Insert Missing Values

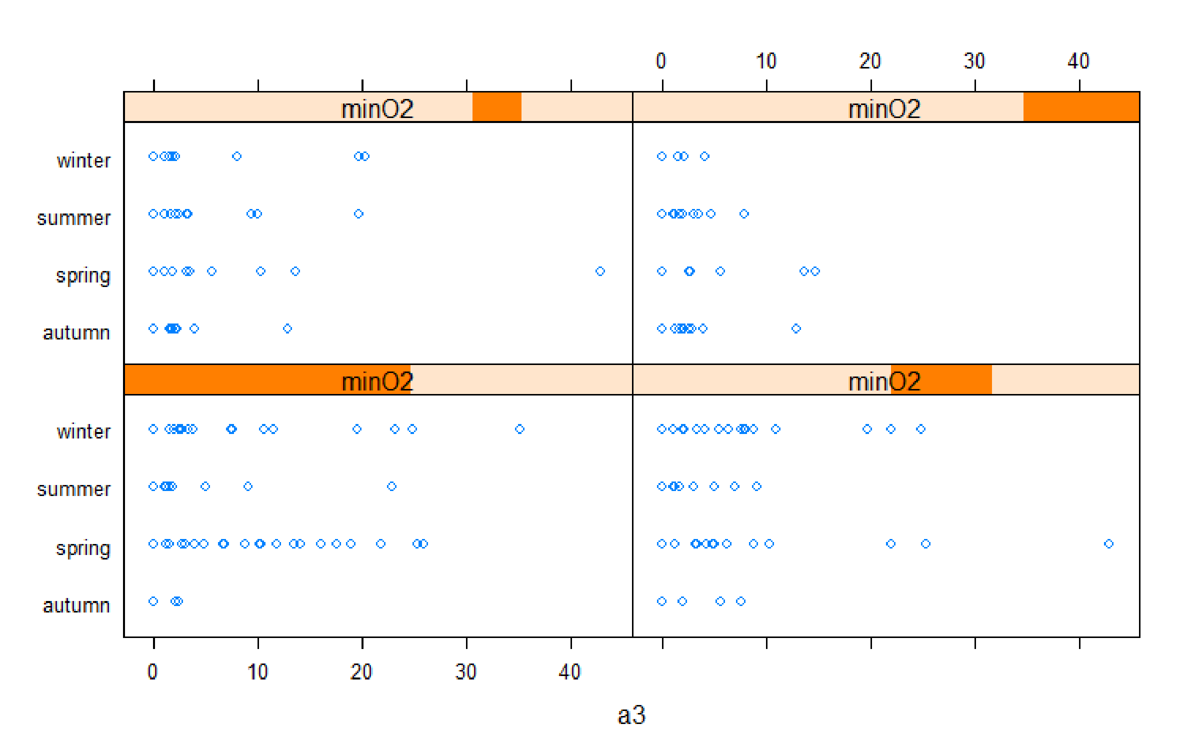
Modeling:
Multiple Regression,
Regression Trees, Random
Forest

Analysis:
Findings, Conclusion,
Next Steps

Exploratory Data Analysis

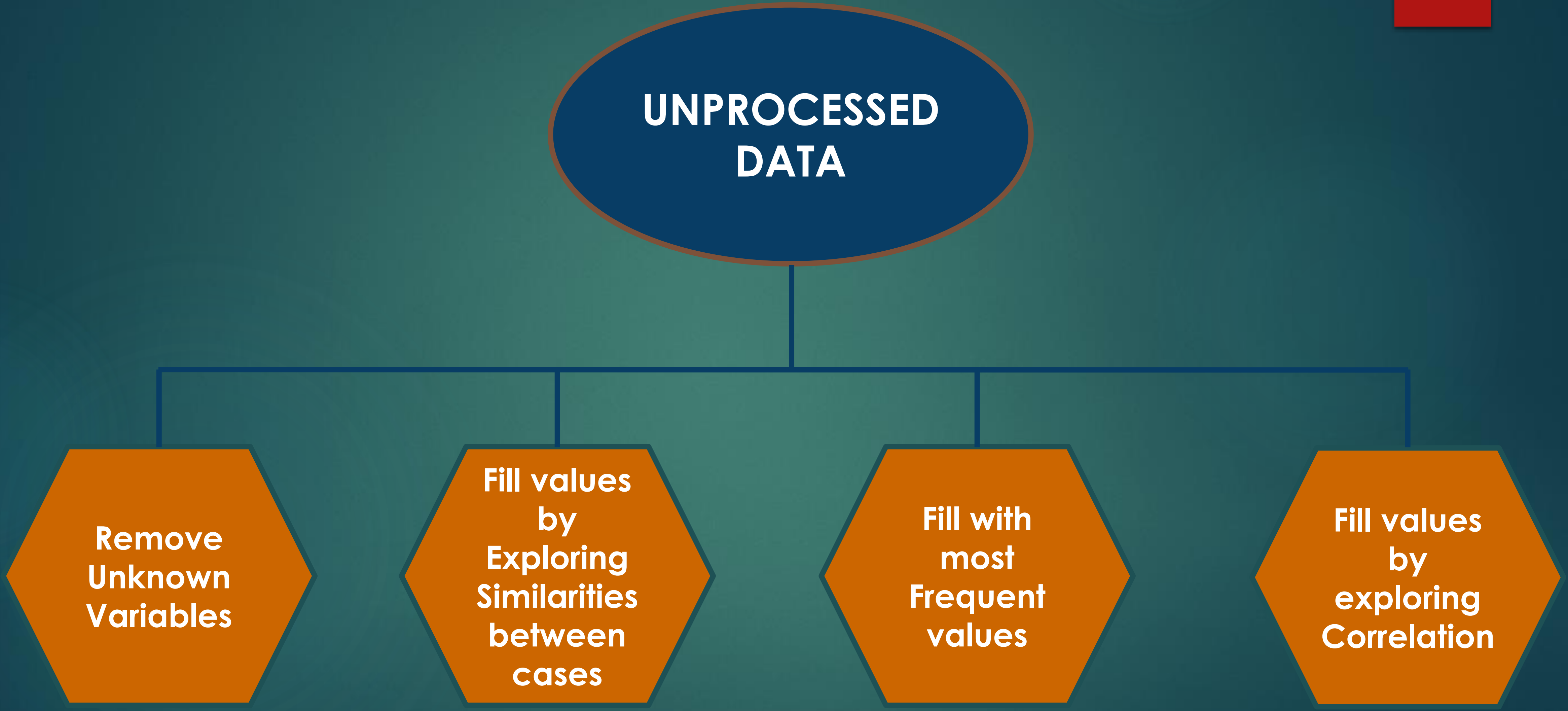


Scatter plot on NH4



stirplot for season vs algae type a3

DATA PREPROCESS



UNPROCESSED DATA

Remove Unknown Variables

Fill values by Exploring Similarities between cases

Fill with most Frequent values

Fill values by exploring Correlation

MODEL-Multiple Regression

- Apply Regression model on the processed data set for algae category a1

- The results show that the variance of model is 32%

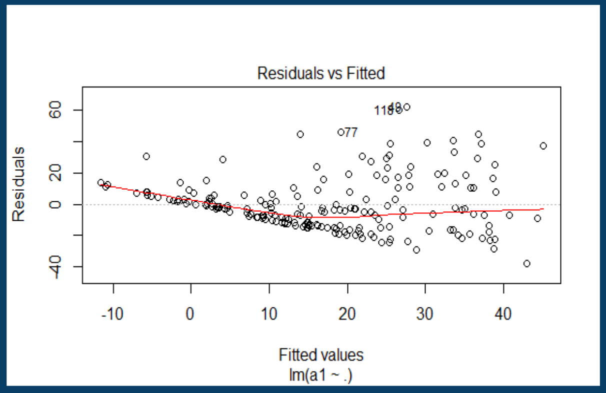
- Removal of Coefficients using Backward Elimination

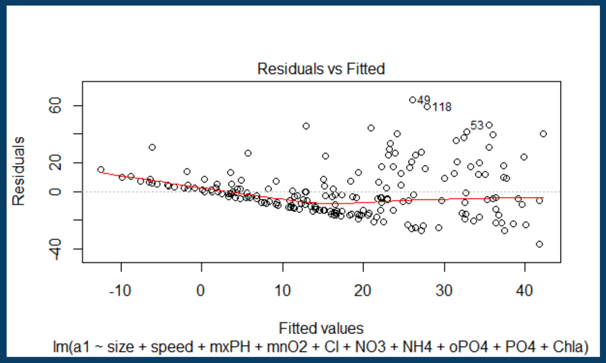
- Fit of the model is improved to 32.8%

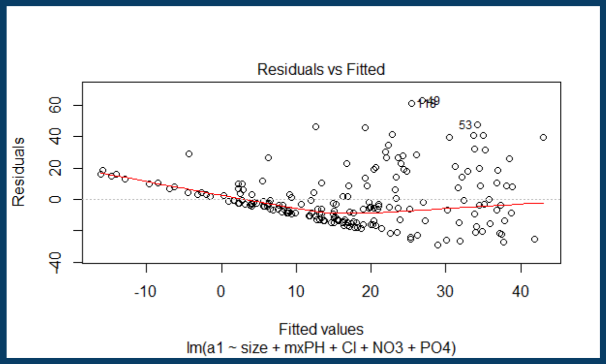
- Comparing two models using Anova

- Step function is applied , proportion of variance explained is 33.24%

- The model tells us that phosphate content in lake stimulates growth of algae at a high rate







Why Multiple Regression?

1

Less complex!
Allows the simulation of different scenarios by varying the values of input variables.

2

Selects the best fitted combination of predictor variables.

3

Variable Insignificant?
Uses backward elimination method to delete!

4

Allows multiple independent/predictor variable to be part of regression model

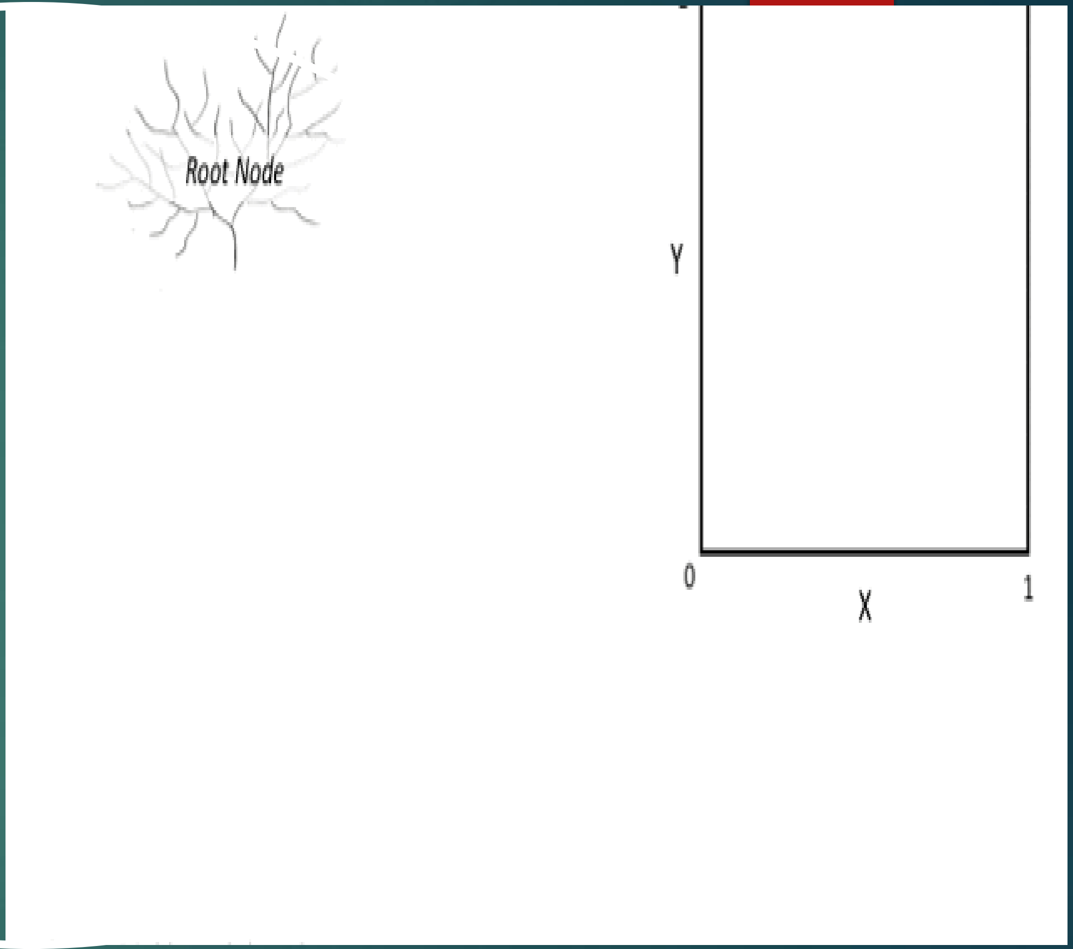
MODEL - Regression Trees

In decision analysis, a regression tree is used to visually and explicitly represent decisions and decision making

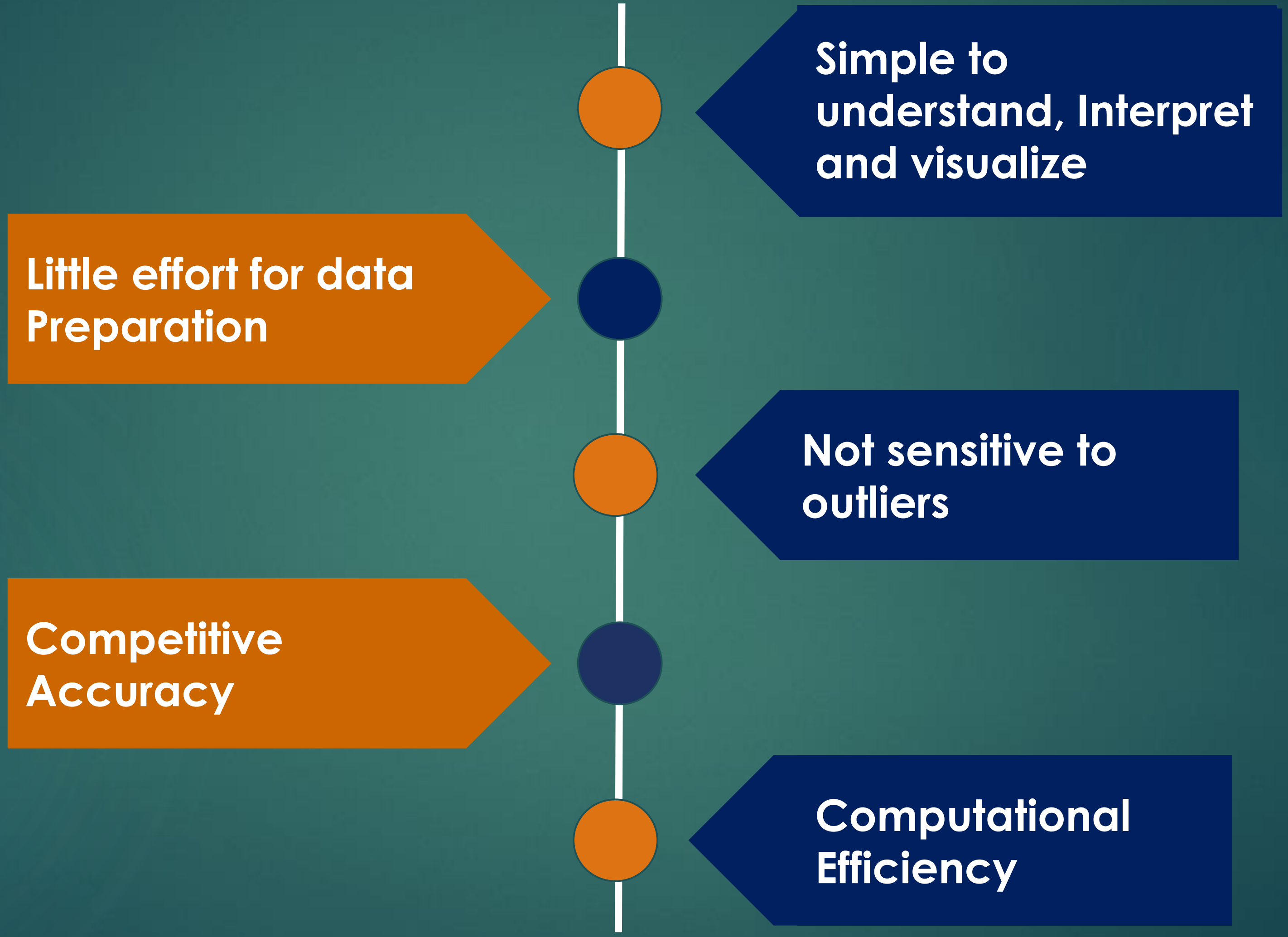
The splitting process results in fully grown trees until the stopping criteria is met

At the leaves we have the predictions of the tree which is a average value for regression task

The fully grown tree is likely to overfit. Pruning technique is used to tackle the same



Why Regression Tree?



Simple to understand, Interpret and visualize

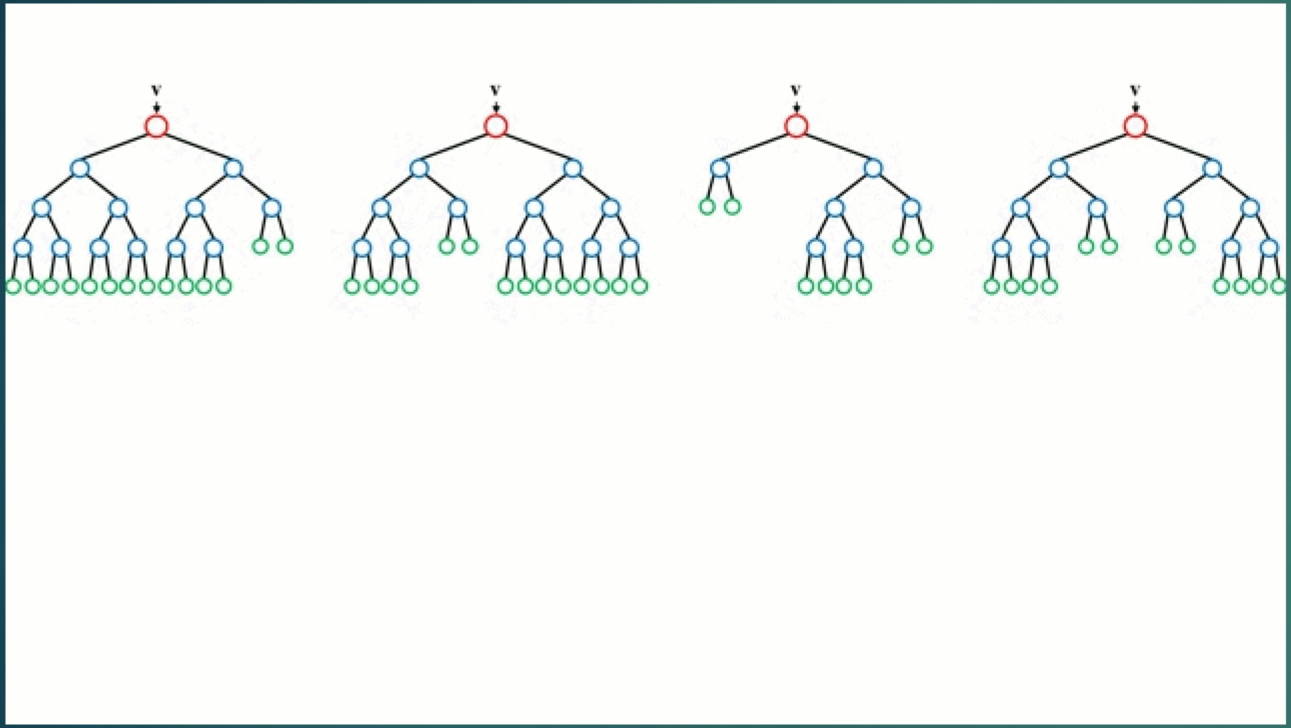
Little effort for data Preparation

Not sensitive to outliers

Competitive Accuracy

Computational Efficiency

MODEL - Random Forest



Random Forest algorithm uses multiple random decision trees.

The cross-validation experiment is performed on three variants of random forests, each with a different number of trees in the ensemble

The forest chooses the average of the outputs of different trees in the case of regression tasks

Random forest does not overfit the model, the best scores confirm the advantages of the ensemble approach.

Why Random Forest?

Generates an internal unbiased estimate of the error as the forest building progresses

Handles large dataset with Higher Dimensionality

One of the most accurate learning algorithms available

It won't overfit the model

Handles the missing values and maintains accuracy for missing data

SUMMARY

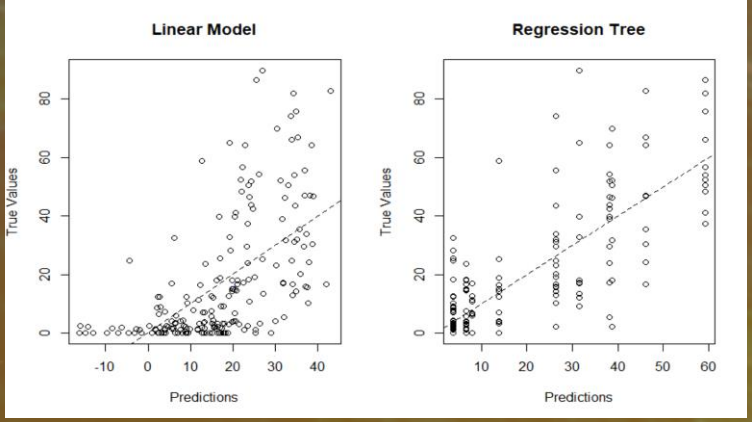
- Harmful algae blooms is an important issue in freshwater lakes and coastal areas particularly lake Erie.

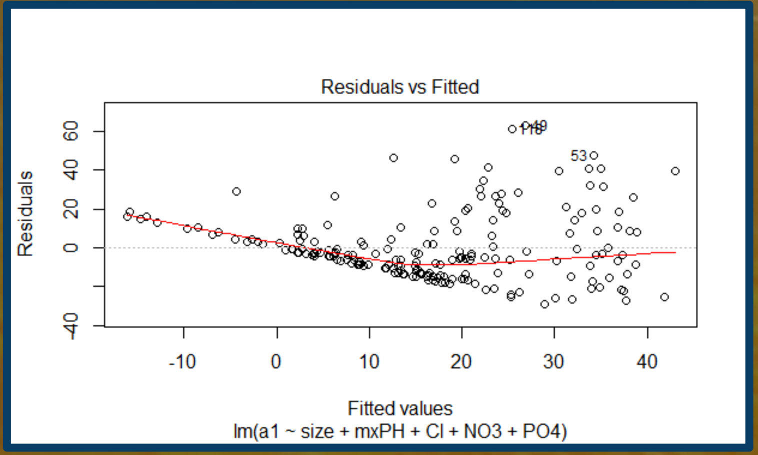
- We applied various models like multiple regression, regression trees and random forest for harmful algae bloom prediction.

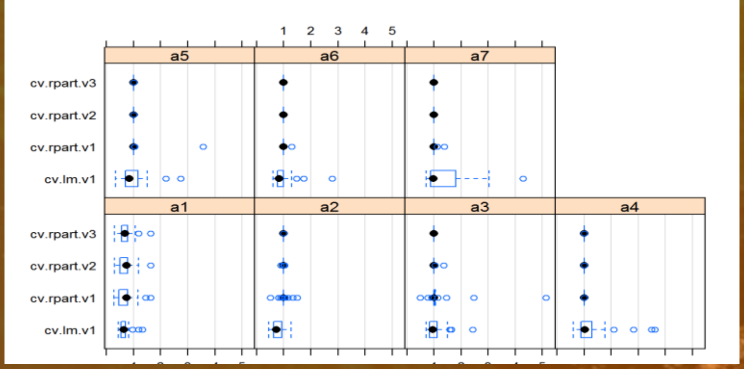
- The Regression Trees worked better than Multiple Regression on the test data. The NMSE value was lower for Regression Trees.

- We applied Multiple Regression to the data set and the efficiency was observed to be 33.24% on one of the type of algae category based on NMSE value.

- Random Forest works best on 4 of the algae types (Output) based on bestscores value







Next Steps

- We plan to investigate further this application problem, trying to overcome the failure of our models in terms of precision in some situations
- We intend to explore other alternatives to current usable by improving data collection and modeling strategies
- We plan to implement artificial neural networks, we selected ANN because it is the most widely used method in water resources variable modelling.
- We also intend to increase the accuracy by using more efficient modelling techniques