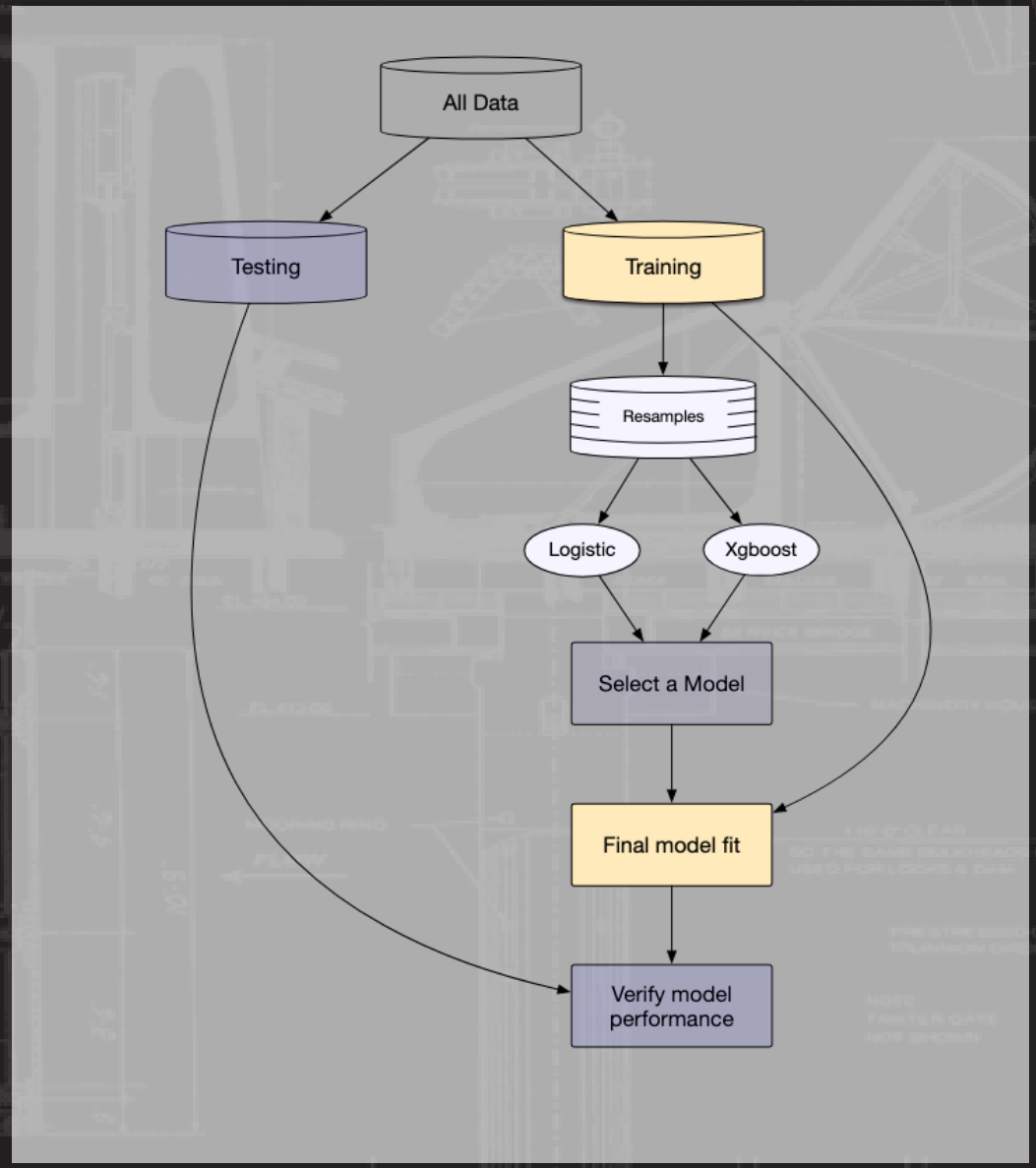
MACHINE LEARNING IN R FOR ARMY CORPS ECOLOGICAL MODELING

Edward Stowe, PhD

Environmental Lab

EMRRP Webinar
January 2026



U.S. ARMY

US Army Corps of Engineers®

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

OVERALL OBJECTIVES

- Demystify machine learning, ML terminology, and Random Forest
- Understand key elements of a good workflow for machine learning (and modeling in general!)
- See how the `tidymodels` package facilitates this kind of workflow
- Understand how this kind of analysis can apply to USACE ecological modeling/planning scenarios

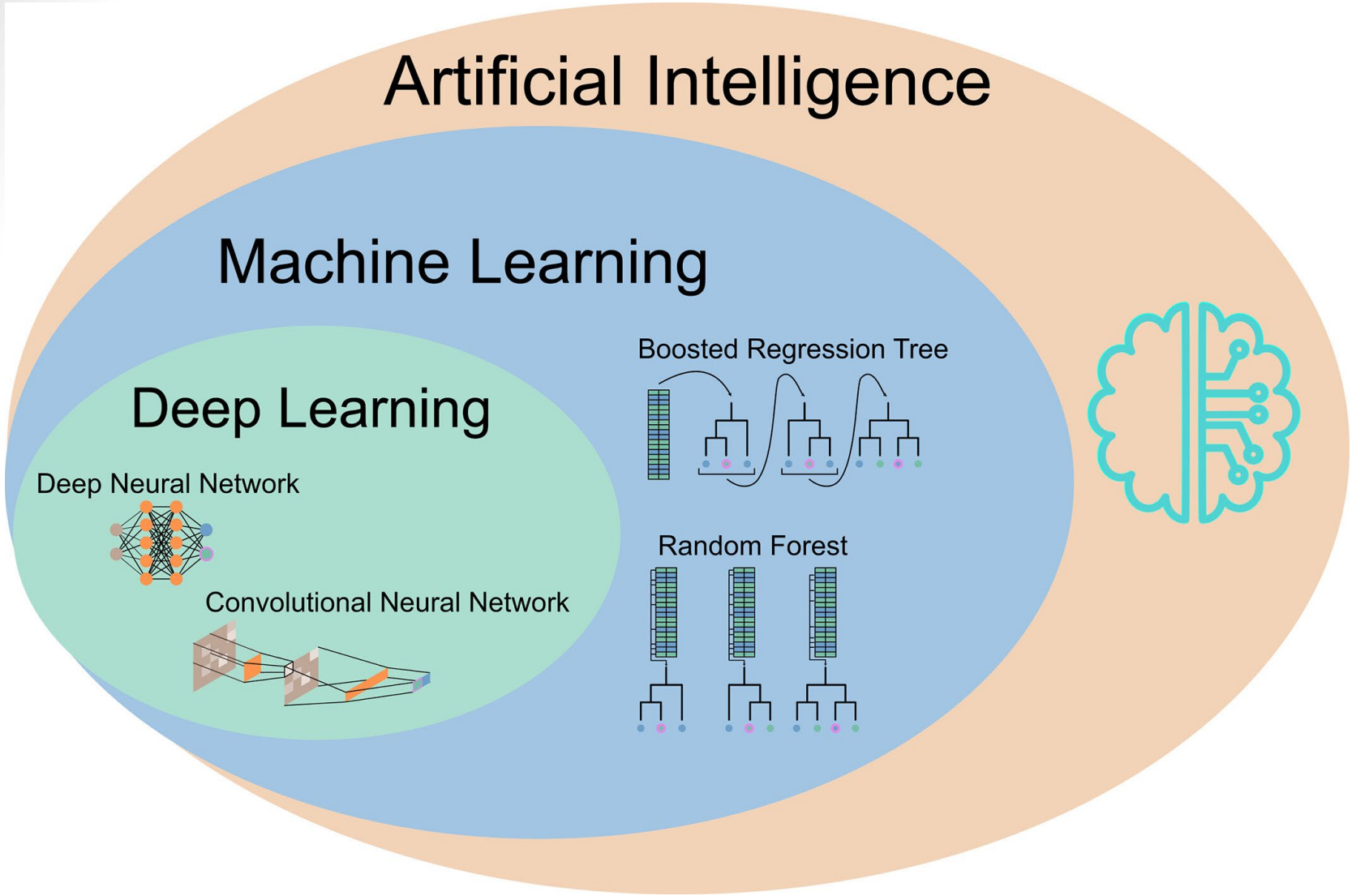
MACHINE LEARNING

Algorithms that learn to recognize patterns in data and can then make predictions in new scenarios

When is it most useful?
- With moderate-to-ample data
- When prediction is the goal

Is it accessible to users?
In many cases, yes. Packages like `tidymodels` in R help.

Aren't ML models black box models?
Not always



Machine learning

- Support vector machines
- Decision trees
- K-means clustering
- Artificial neural networks (simple)

Deep learning

- Various neural networks
- Large language models (LLMs)

Pichler and Hartig 2023

SOME IMPORTANT TERMS

Supervised models: models with a response variable
Unsupervised models: no response variable (e.g., clustering, dimension-reduction)

Classification: predicting a category (presence/absence; classifying an image)
Regression: predicting a number/continuous variable

Parameters**:** predictor variables in a model (e.g., depth, dissolved oxygen); sometimes called features
Hyperparameters**:** model settings that control the ML process (number of trees; minimum n)

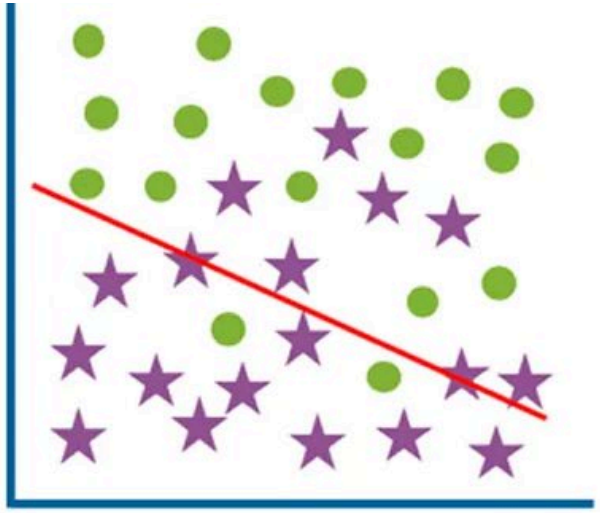
Training: teaching the model how to make predictions
Testing: assessing how well the model has learned

MODEL COMPLEXITY

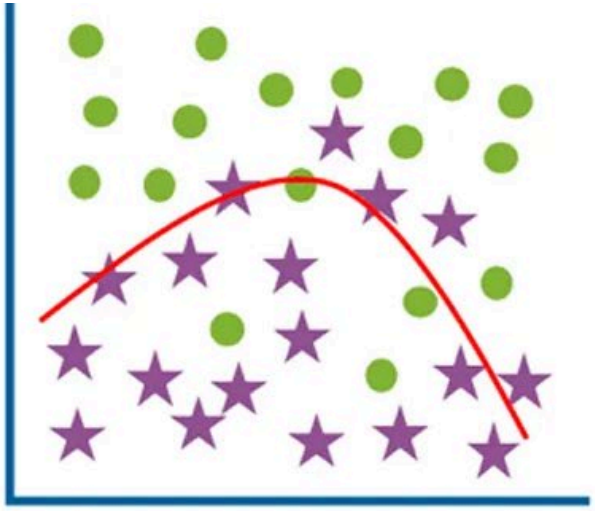
Fundamental tradeoff

More variables = **Better** model performance but **Worse** at predicting new data

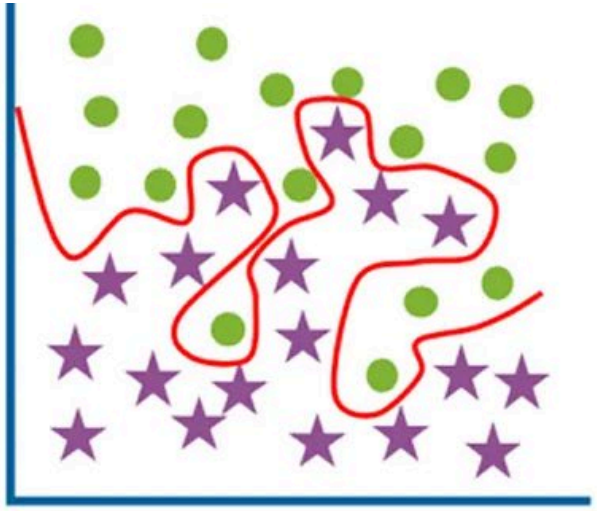
Underfit



Appropriate



Overfit



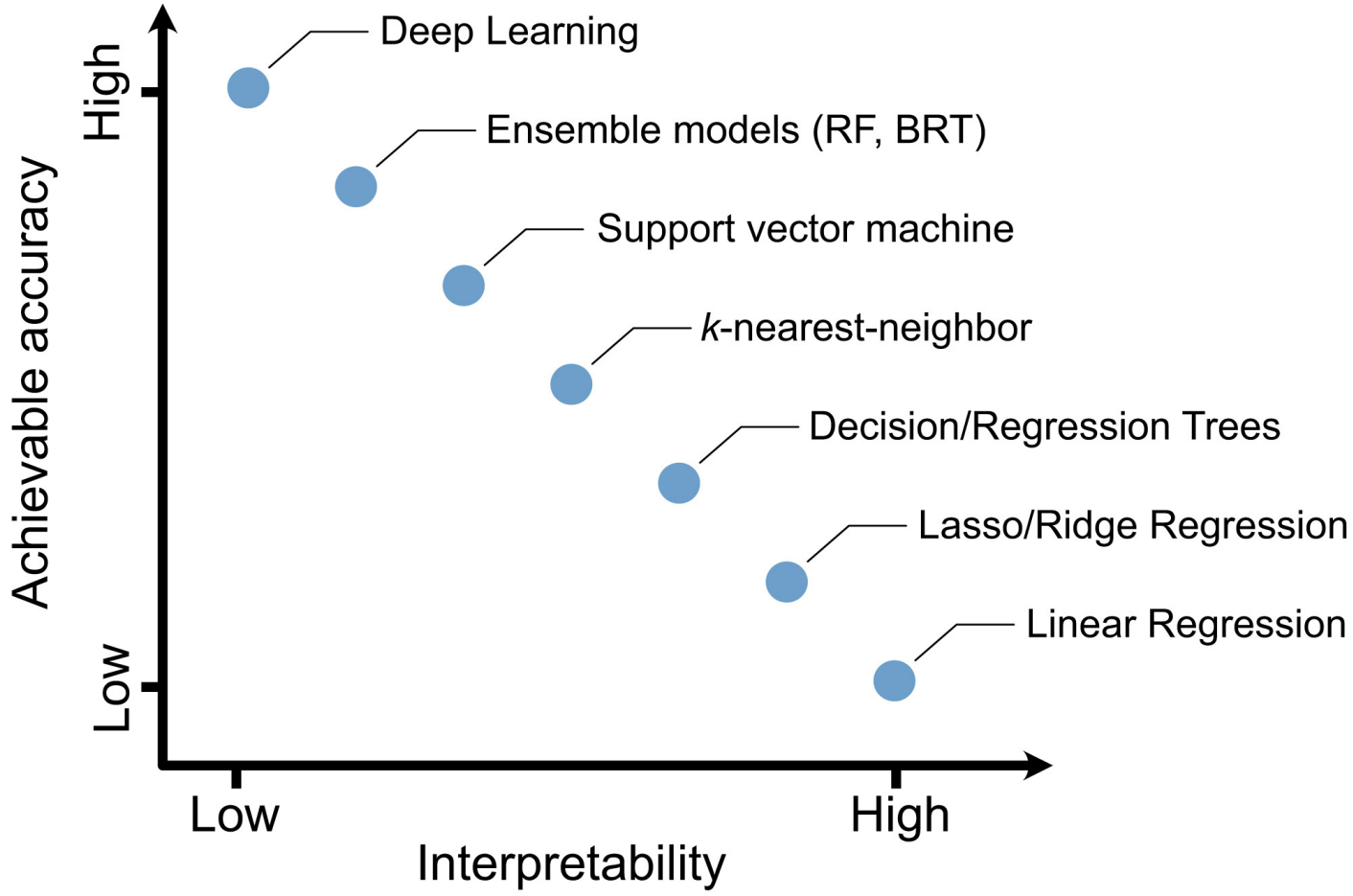
Intermediate complexity = optimal compromise

WHAT ABOUT THE BLACK BOX?

Some models achieve high predictive accuracy but are hard to understand

This can have consequence:

- Harder to communicate
- Lower trust
- Fewer insights into processes

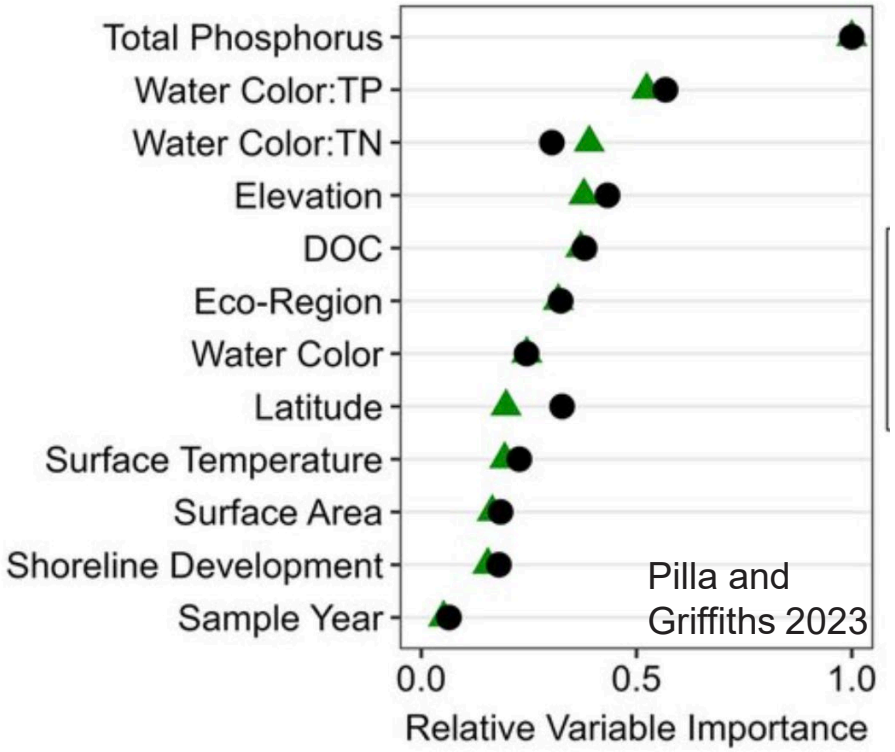


Pichler and Hartig 2023

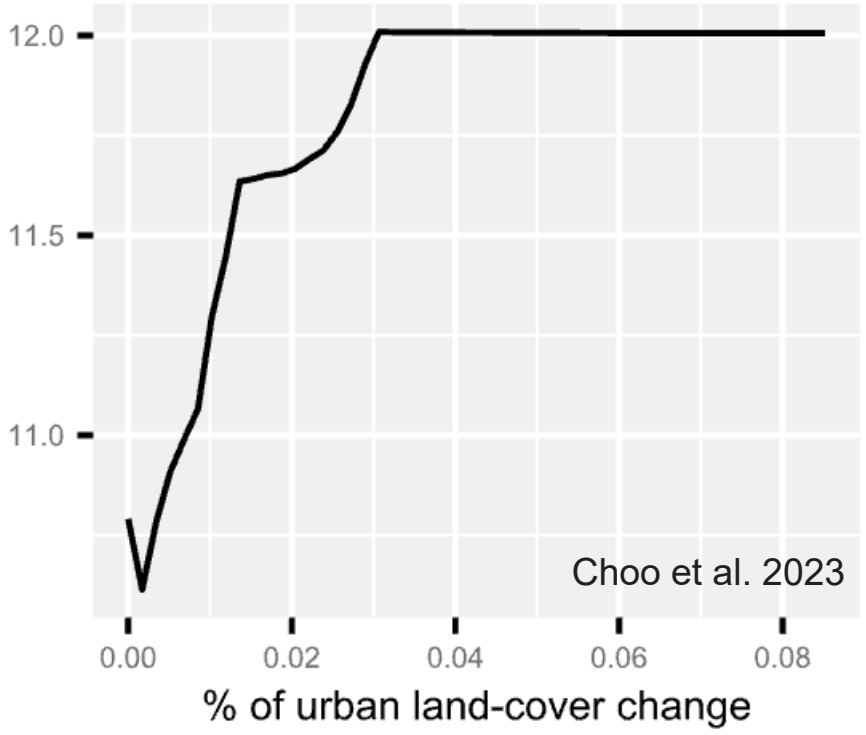
WHAT ABOUT THE BLACK BOX?

Explainable AI approaches can help.

Variable importance



Partial dependence plots



Other approaches:

- Shapley values
- Perturbation results

USACE PLANNING APPLICATIONS

- Assessing relationships between environmental data and species/ecosystem attributes
- Validating/updating HSI models
- Identifying design criteria/thresholds
- Assessing species associations

Example:
Ecological modeling for planning
– Input: 4 – 8 variables
– Output: Probability of occurrence of desired taxa
– Probabilities can be converted to suitability scores

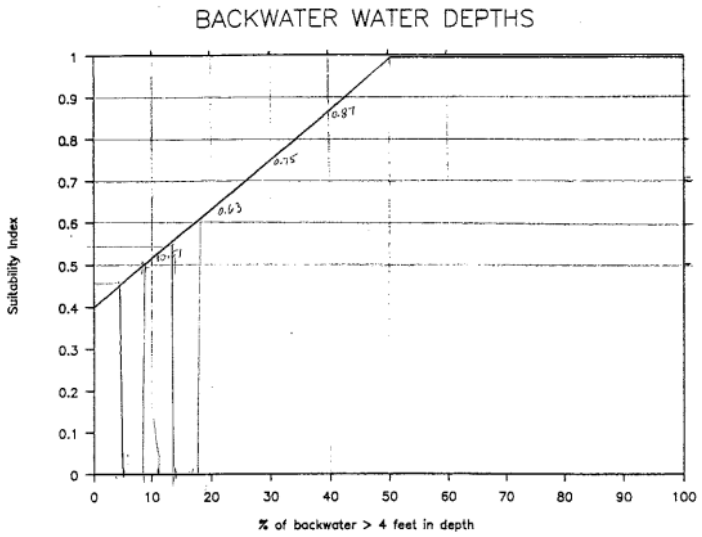
CASE STUDY – UPPER MISSISSIPPI RIVER LENTIC FISH

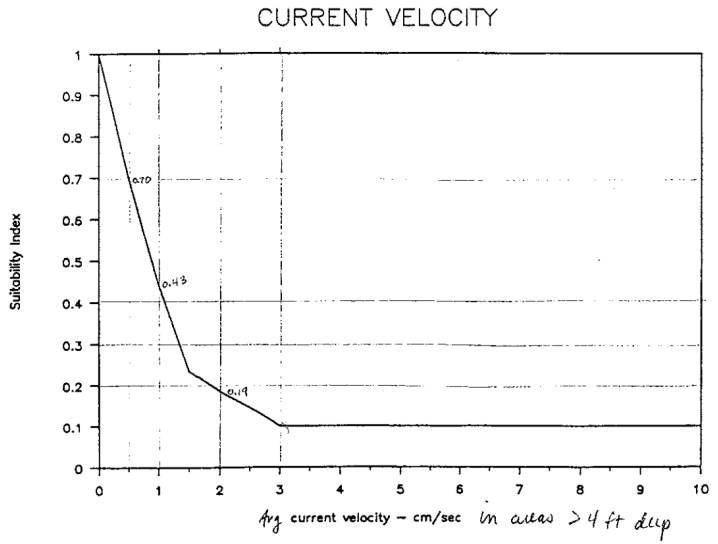
Restoration efforts in the UMR target conditions for overwintering lentic fish

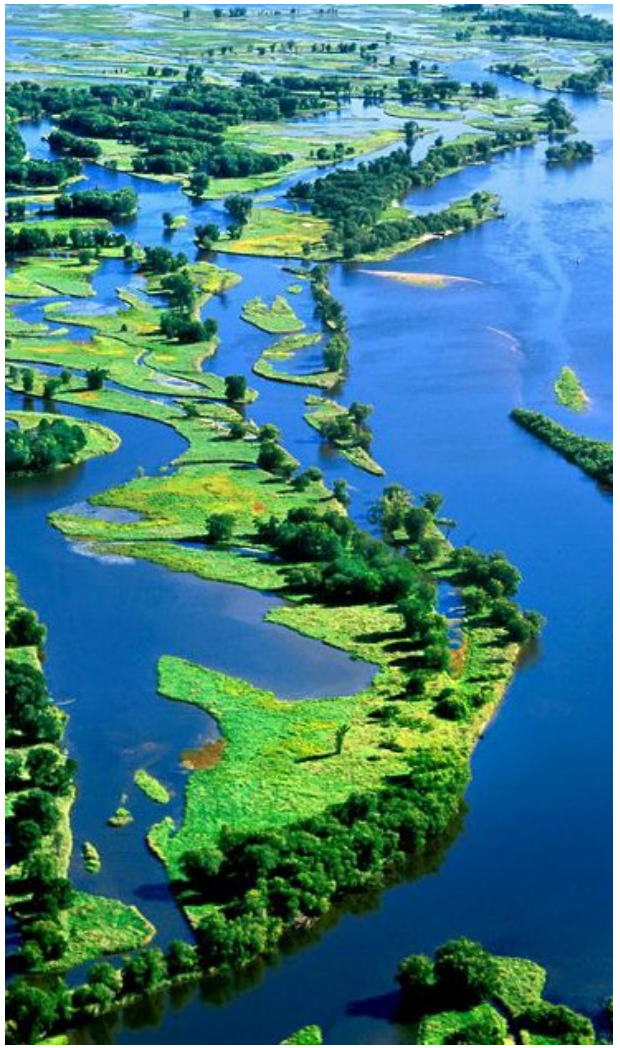
Simple Bluegill HSI model used to predict effects

Does the wintering HSI model apply to summer?

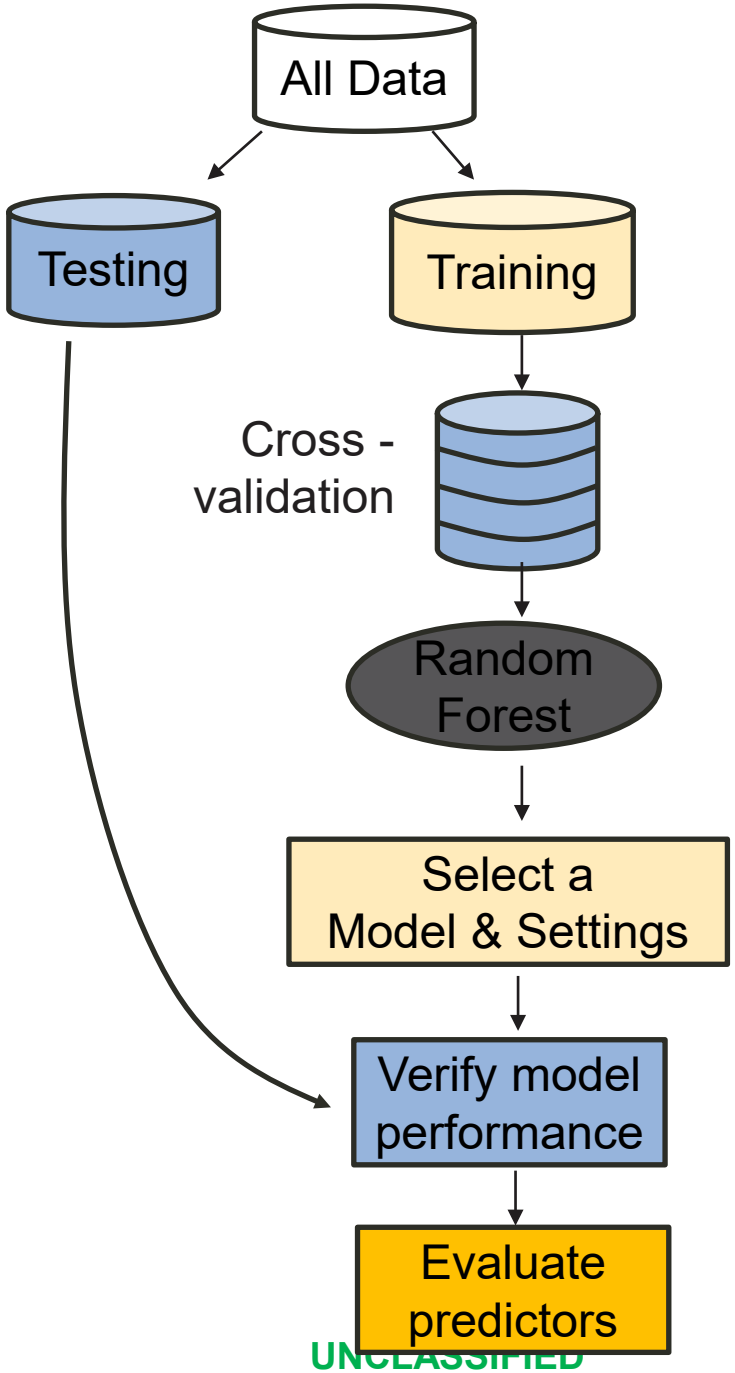
Are bluegill design criteria supported by the data?







MACHINE LEARNING WORKFLOW



Data splitting

Model Tuning

Model Assessment/Selection

Model testing

TIDYMODELS PACKAGE

An integrated set of R packages that can be used in comprehensive modeling/machine learning workflows.

Streamlines many tedious aspects of machine learning, like pre-processing data, using different algorithms, data splitting and cross validation, and a lot more…

DECISION TREES

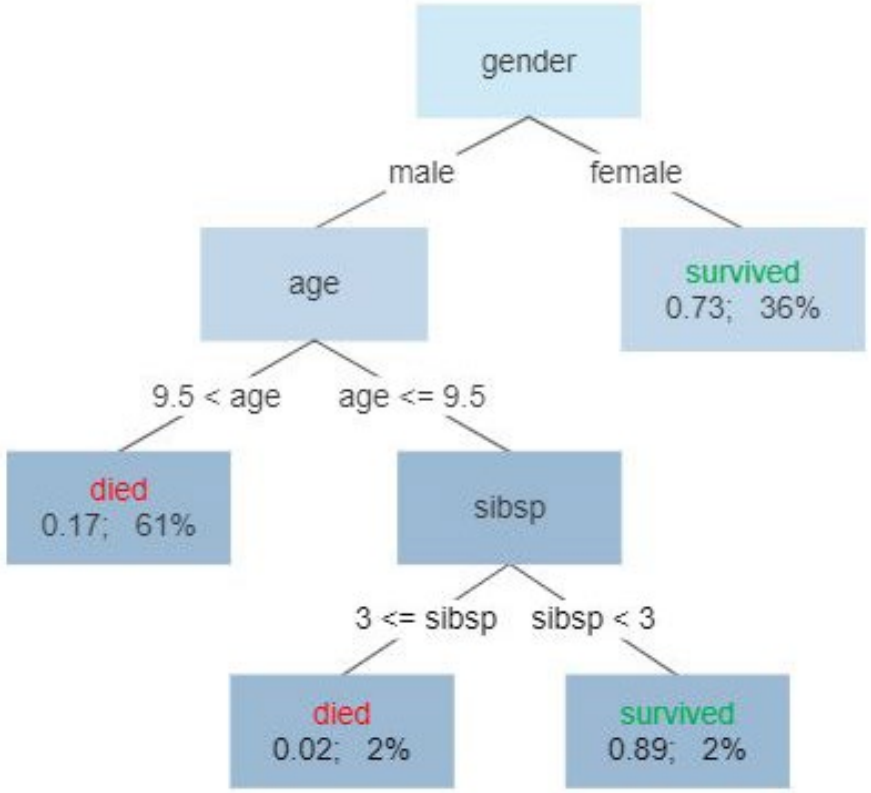
Goal: predict a value of a target variable based on inputs

Model finds decision nodes that help predict the outcome

Interpretable, but tend to **overfit**

Recap: overfitting means learning the training data too well and making poor predictions in novel scenarios

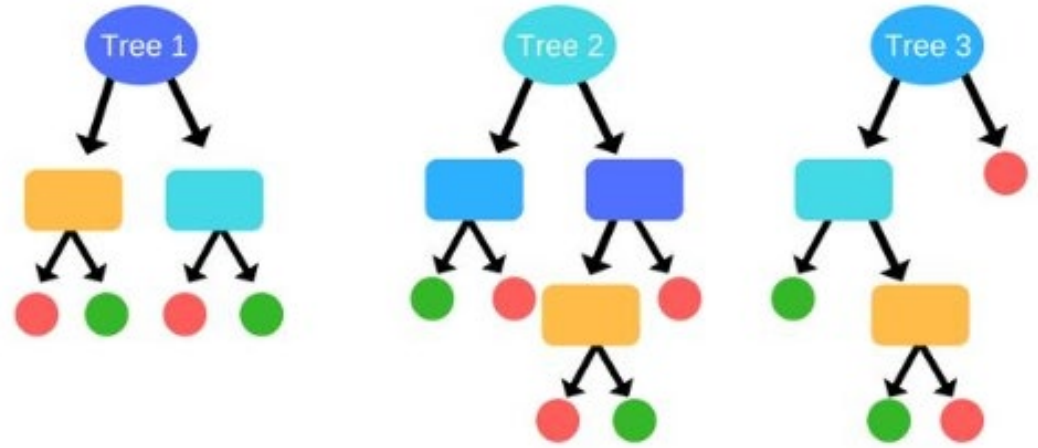
Survival of passengers on the Titanic



Credit: gilgoldm

RANDOM FOREST

- Many trees (e.g., 1,000 - 2,000)
- Each tree:
  - Subset of the data (rows)
  - Subset of the predictors
- Each tree votes on the likely outcome
- “Ensemble” scheme avoids overfitting



Credit: dataaspirant.com

UNCLASSIFIED

CASE STUDY OBJECTIVES

Predict bluegill occurrence in Upper Mississippi backwater habitats as a function of environmental predictors & evaluate relationship between bluegill and some environmental predictors

Attendees can observe:

- Ease of data splitting and cross validation
- How to pre-process data and select a model/algorithm
- How to tune and assess model performance
- How to visualize some of the black-box elements of an ML model
- How these analyses can be used to alter/update USACE ecological models

UNCLASSIFIED

ACKNOWLEDGEMENTS

Research funding

- Ecosystem Management and Restoration Research Program (EMRRP)

Feedback

Survey form at the end of the session

Contact info

Ed Stowe, PhD

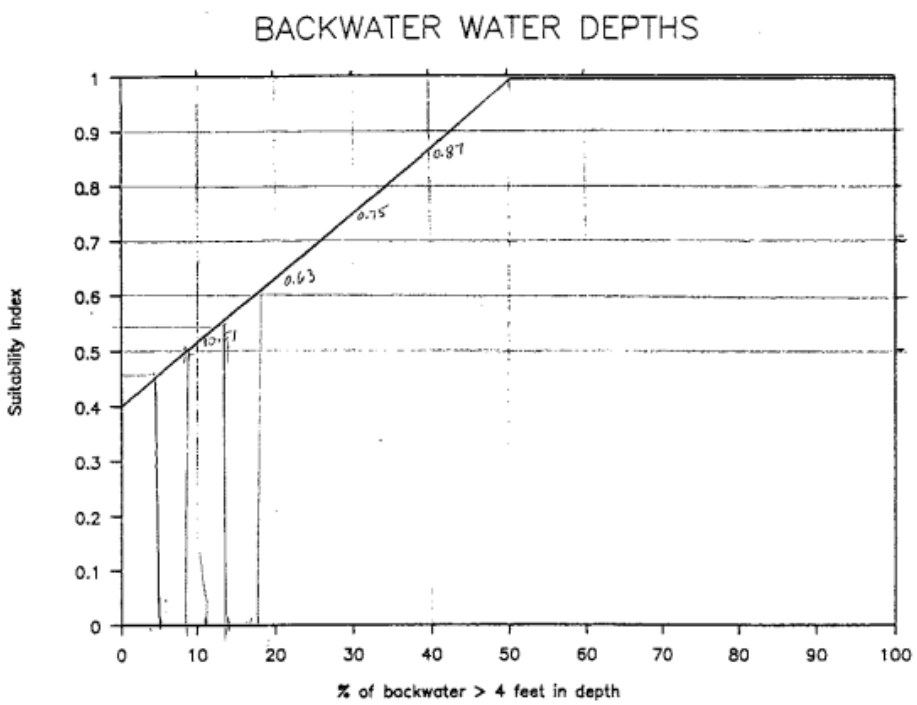
Athens, GA

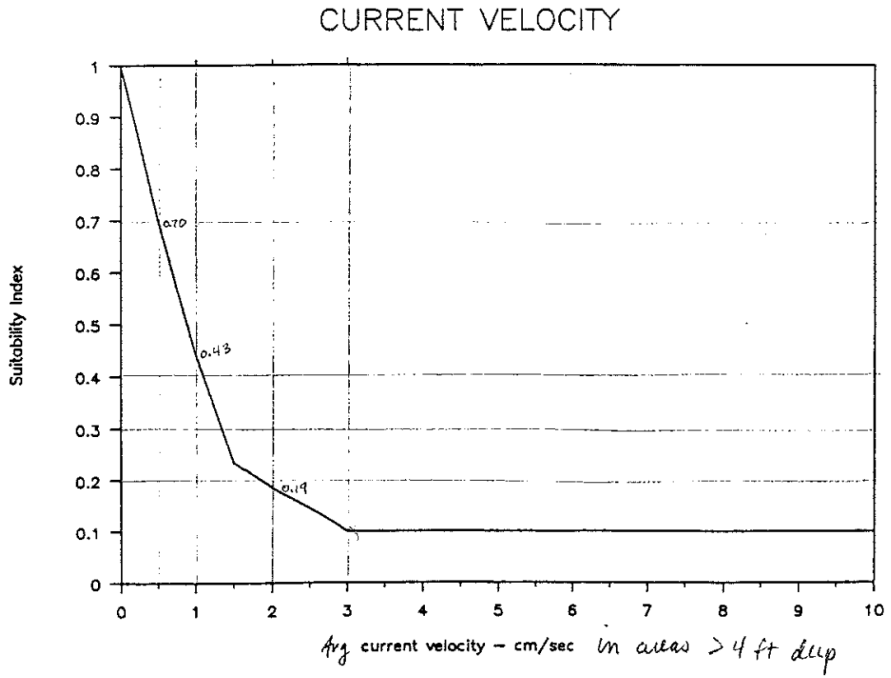
Edward.S.Stowe@erdc.dren.mil

Current web location

https://usace-wrises.github.io/USACE.EcoMod.Training

UNCLASSIFIED





UNCLASSIFIED

TAKEAWAY MESSAGES

Streamlines tedious aspects of modeling (e.g., different syntaxes, data splitting; cross-validation)

Machine learning can be used to transform models used for planning

Many of these elements should ideally be incorporated in ecological modeling more broadly
– Data-splitting
– Testing on out of sample data
– General modeling workflow
– Minimizing bias introduced by pre-processing

SURVEY FORM
https://forms.office.com/r/AARtQjXKdZ