



# Biointerphase – The Bat-Detection Model AI Studio Presentation

BTTAI BOS  
Nov 10, 2023



# Introductions



# Meet Our Team!



**Anders Freeman**  
Wellesley College



**Anusha Bandaru**  
Northeastern University



**Iris Yang**  
Tufts University



**Linda Dominguez**  
Wellesley College



**Yuhan Wang**  
Smith College



# Our AI Studio TA and Challenge Advisors



**Leandra Marie Tejedor**  
AI Studio TA



**Noah Snyder**  
Challenge Advisor



# Presentation Agenda

1. AI Project Overview ~ 5 min
2. Data Understanding & Data Preparation ~ 5 min
3. Modeling & Evaluation ~ 10 min
4. Final Thoughts



# AI Studio Project Overview

# Background

- White Nose Syndrome(WNS) is a fungal disease that infects skin of the muzzle, ears, and wings of hibernating bats
  - WNS bats during hibernation are woken up, causing them to use imperative fat reserves and often leading to starvation and death.
- Impact:
  - Has ravaged North American bat populations.
  - Bats are valuable contributors of pollination.
  - Bats contribute to pest control efforts

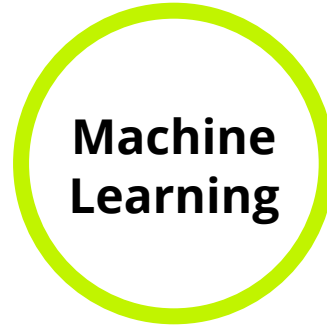




# Our Goal



Understand the outlook  
for bat population decline  
in North America due to  
WNS



Create Machine Learning  
models to predict what  
features signal population  
decline



Provide insight to guide  
efficient bioengineered  
solutions to combat WNS





# Business Impact

- Inform Biointerphase of the most efficient action to combat WNS
  - current situation of WNS
  - locate key regions/ most endangered population
  - resource prioritization
- Provide basis for future work on combating WNS
  - availability of public datasets on WNS
  - preliminary ML models on WNS-related data
  - possible research directions



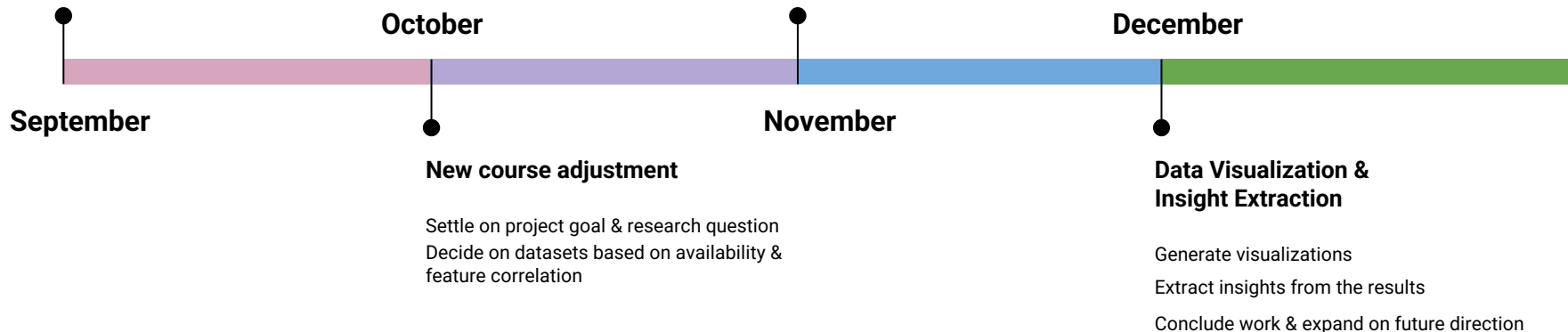
# Our Approach

## Problem Understanding & Data Exploration

Research WNS on official websites  
Explore past researches on similar topics  
Gather & explore datas

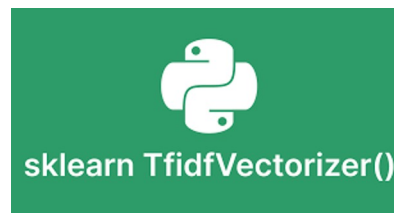
## Two-Track Modeling

Split into 2 groups to build ML models parallelly  
Exchange insights between groups  
Troubleshoot/ debug collaboratively





# Resources We Leveraged





# Data Understanding & Data Preparation



# Finding Data

- At the beginning, we were still understanding the scope of our problem and deciding on whether to use non-acoustical data or not to see how much decline a certain bat population had.
- We found out that non-acoustical data was going to be very timely to train, and instead of predicting the amount of decline in a certain bat population, we were going to predict if a bat population in a certain location had white nose syndrome or not.



# Our Data

- We combined 3 datasets into one to analyze.
- Two datasets are composed of what specific type of fungi a bat population has.
- The other is composed of determining whether a population has WNS or not given the location.
- Dataset has columns for:
  - Location
  - Complete biological classification
  - Concentration of fungi
  - Time of sample collection
  - If WNS had been present in the population at that time
- The dataset overall has 800 entries



# Fungus Classification Snippet

Fungus Classification – Phylum	Fungus Classification – Class	Fungus Classification – Order	Fungus Classification – Family	Fungus Classification – Genus	Fungus Classification – Species
Ascomycota	Eurotiomycetes	Onygenales	Gymnoascaceae	NaN	NaN
Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	Debaryomyces	hansenii
Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	Debaryomyces	hansenii
Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	Debaryomyces	NaN
Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	Debaryomyces	NaN



# Modeling & Evaluation





# Modeling and Variations

We have used 2 modeling tracks:

- Track 1 - NLP
  - Took all features and put them as a natural text “sentence” and fed that data into a natural language processing pipeline.
  - Created the actual neural network with the keras library, while training it on the features in the set
  - Finally, we plotted the training and validation loss and accuracy.
- Track 2 - Random Forest Model
  - Transformed all the features into single categories and concatenated them into one dataframe.



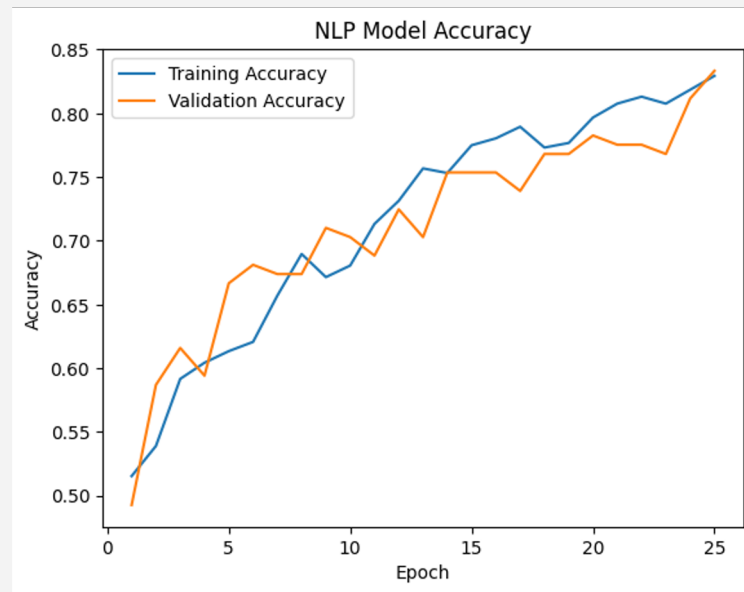
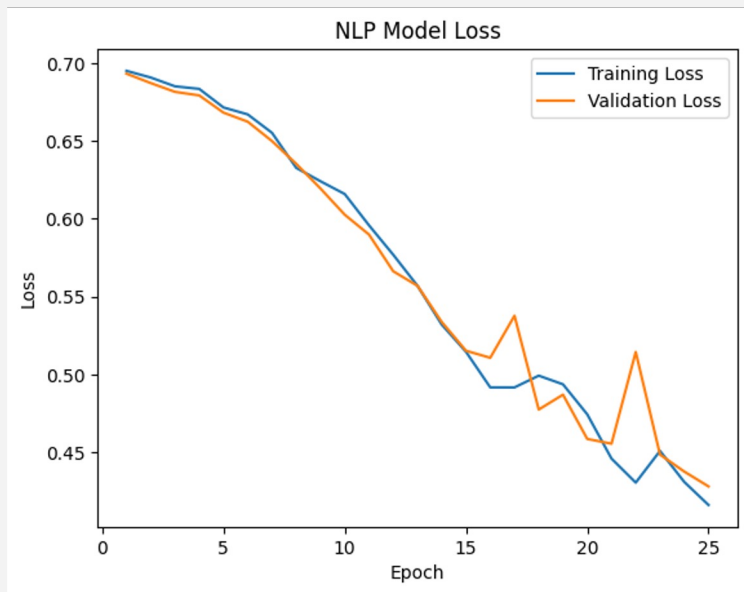
# NLP Model - Overview

- Natural Language Processing
  - Neural Network — ML models inspired by the structure and functioning of human brains
  - Specific to understand, interpret, and generate human language by processing textual data
    - e.g. ChatGPT, Translation, Siri, Grammarly
- Our NLP Model
  - Feature: a giant string with all features combined as the text input
  - Label: 0 – WNS is NOT present in the population; 1 – WNS is present in the population
  - Reason
    - 9 out of 11 features are strings
    - Fungus classification on different levels are names
  - Example
    - “2015-01-15 unassigned Hamilton New York 12 Ascomycota Leotiomyces Thelebolales Pseudeurotiaceae Pseudogymnoascus destructans”




# NLP Model - Visualization

Loss & Accuracy of the NLP Model after every epoch



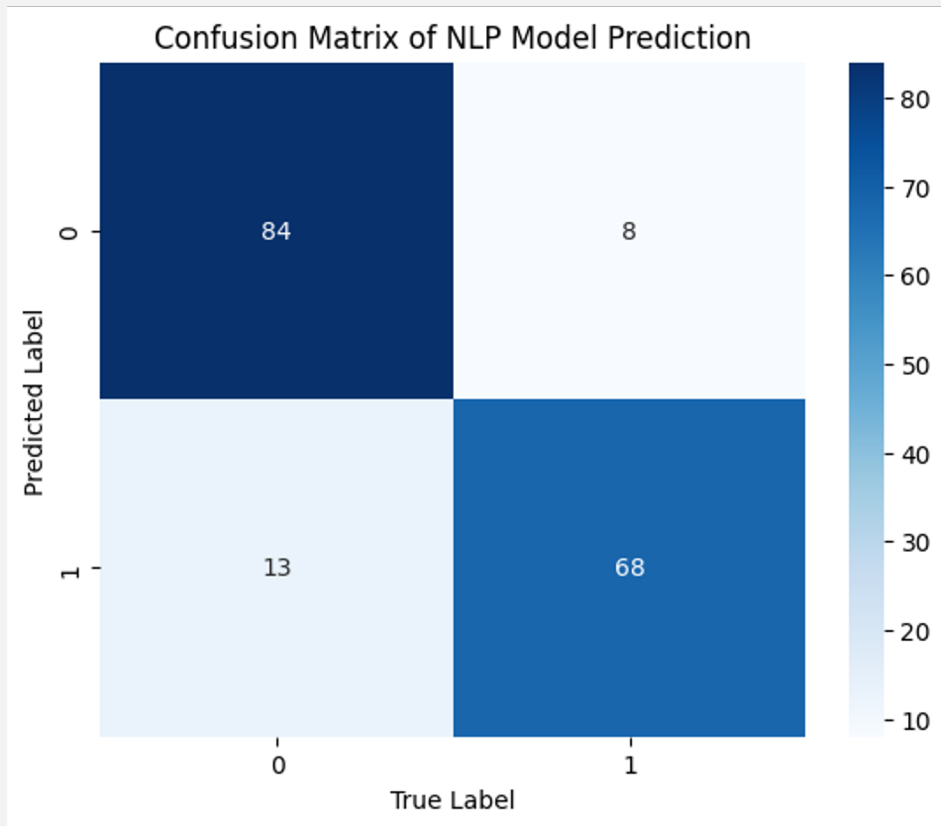


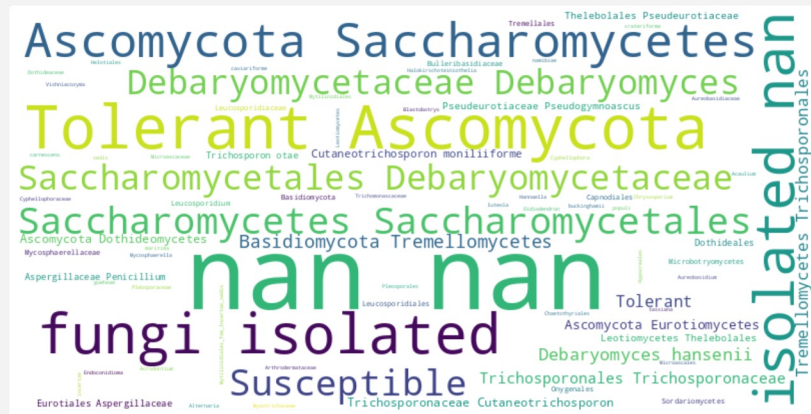
# NLP - Results - Features

Data Included	Model Accuracy
State, County, Date, Host Group, CFU, Fungal Classification	98.26%
Date:	95.95%
State, County:	100.00%
Fungus Classifications:	71.01%
CFU:	83.24%
Host Group, CFU, Fungus Classification:	86.13% 



# NLP Model - Visualizations - Confusion Matrix







# NLP Model - Insights and Key Findings

- The presence of WNS fungus in the population is not directly linked with WNS itself
- Species of fungus are associated with populations that have had WNS in the past



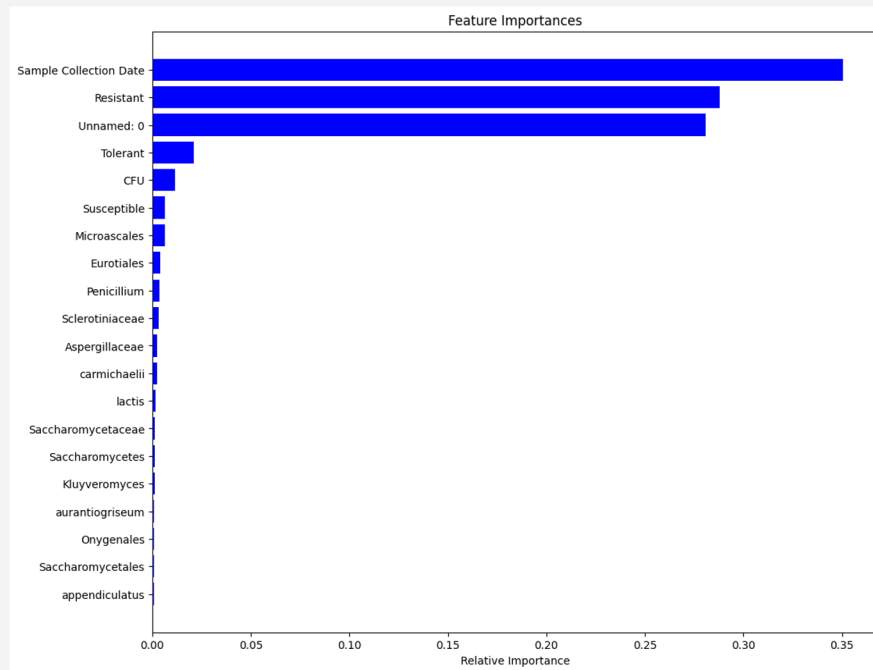
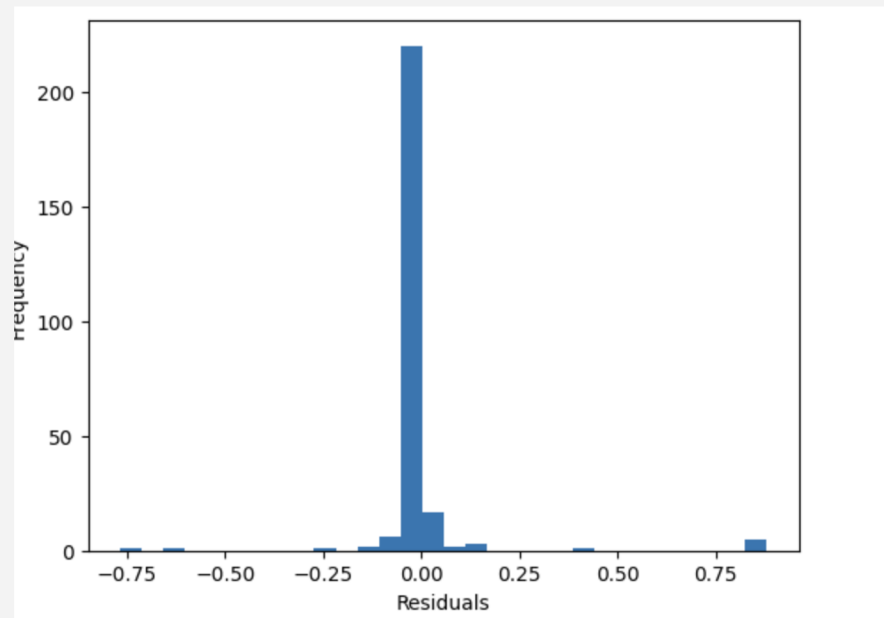
# Random Forest Model - Overview

- A Random Forest model is an ensemble method, meaning it is made up of smaller models. In our case, random forests are made up of smaller decision trees, or estimators.
- Because we found in our NLP model that the columns 'Sample Location - county' and 'Sample Location - state' are highly correlated with whether the population has White Nose Syndrome or not, we removed those columns to account for bias.
- Our Random Forest model:
  - Is more interpretable than a NLP model
  - Splits the data into training and testing sets, as well as a random state to ensure reproducibility
  - Has a max\_depth of 32 and a maximum number of estimators at 300.
  - After the model is fitted and predictions are made, we use the root mean squared error (RMSE) and the R2 score to get our results. The RMSE measures the **difference between a model's predicted values and the actual values**, while the R2 score is the accuracy.





# Random Forest Model - Visualization





# Random Forest - Results - Features

Data Included	Model Accuracy
Base accuracy	92.1%
Removing Sample Collection Date	86.7%
Removing Resistant	97.3%
Removing Unnamed: 0	90.4%
Removing both Sample Collection Date and Resistant	94.9%
Removing both Sample Collection Date and Unnamed: 0	33.6%



# Random Forest Model - Insights and Key Findings

- After performing a grid search with cross-validation to tune the hyperparameters, we found that the best-performing combination was a max depth of 30 and n-estimators of 100.
- We also found that the features that were most correlated with the data were the date and how resistant the population was.
- Sample Collection Date, and to a lesser extent, Unnamed: 0, is crucial to the accuracy of the Random Forest Model. Without it, the accuracy drops by a lot.



# Model Comparison

Model Name	Description	Results	Pros	Cons
Neural Network with NLP	Treating all the features as part of the same text input string and performing binary classification	Accuracy: 0.849	Fast and easy implementation; High Accuracy	Unclear about its mechanism; High reliance on personal interpretation
Random Forest	Using multiple decision trees to gain an overall result for regression.	Accuracy: 0.921	Easier to understand which features are most crucial.	Model is inconsistent; will not necessarily perform well in practice



Final Thoughts



# What We Learned

- The importance of having a clear outline of our research question
- Prioritizing data exploration, analysis, and experiment
- Communicating when we get stuck and working together to figure out answers



# Potential Next Steps

- NLP model
  - Improvements to data cleaning pipeline and time series labeling
    - create time-specific WNS presence label
    - implement different vectorizers
  - Analyze the NLP Model to find correlations between Fungus species and WNS presence
    - evaluating if fungus species is correlated with bat species, and if bat species is correlated with WNS
- Random Forest model
  - Feature engineering
  - Looking into stacking ensemble models
  - Expand our dataset for a more comprehensive model
- Compare 2 models' results for more insights

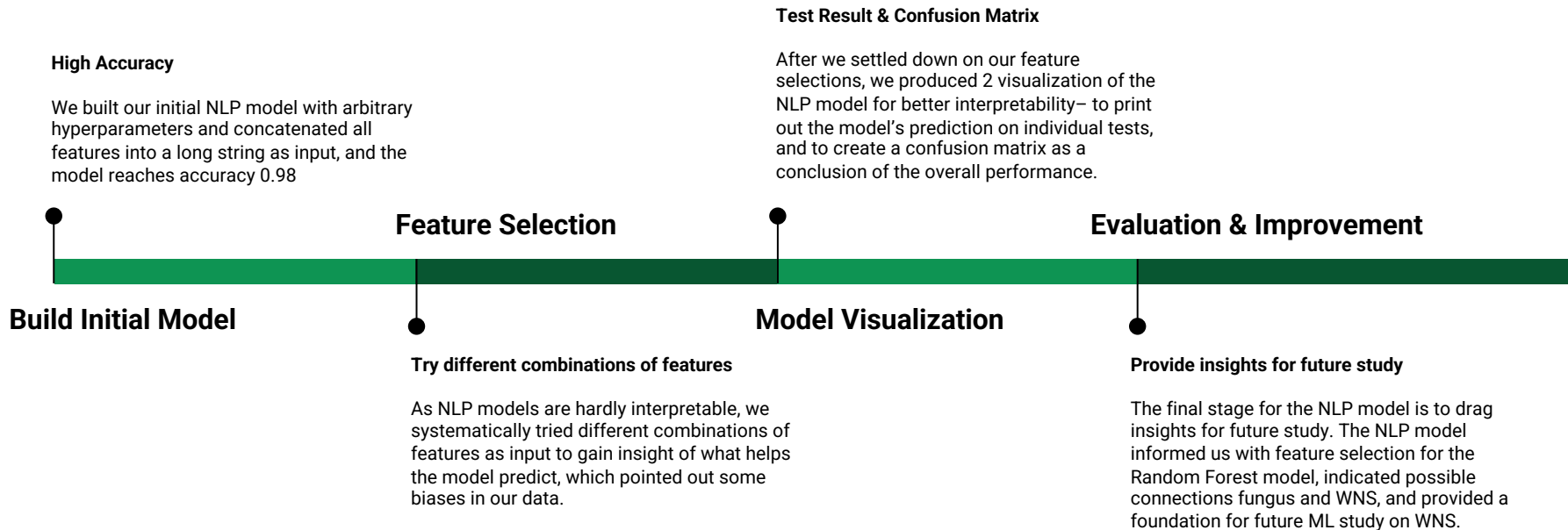


**Thank you!**  
**Questions?**





# NLP Model - Progress





# Random Forest Model - Progress

