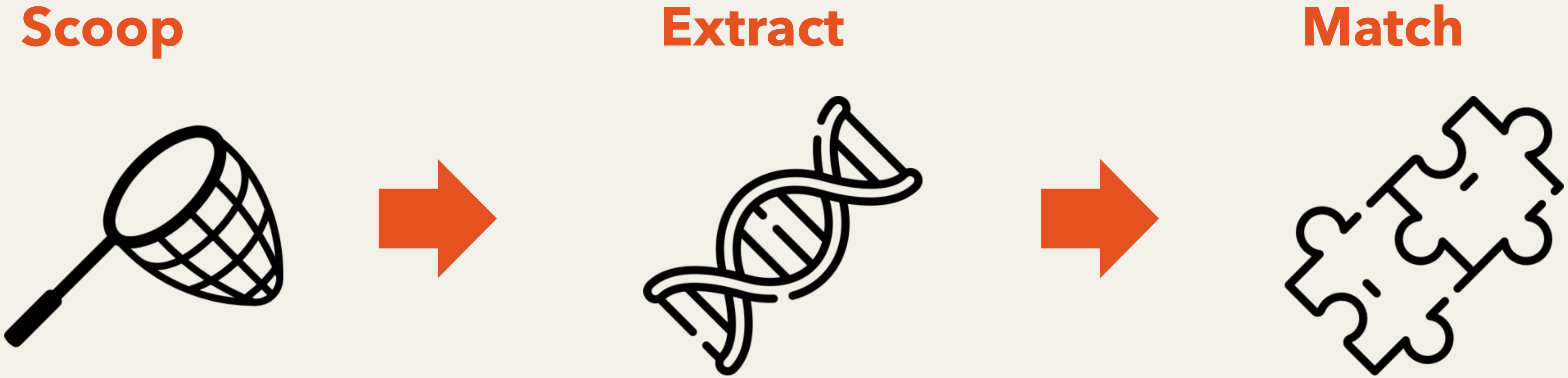
Validating Ocean eDNA Samples Using a Bayesian Probability Model

Group 16: Divya Kalidindi (24475181) | Nate Reed (24110024) | Nitish Rungta (23939001) | Ray Stokes (24423873) | Sweta Manjaly (24400046) | Vinayak Jayananth (24207881)

The eDNA process



Scoop

Extract

Match

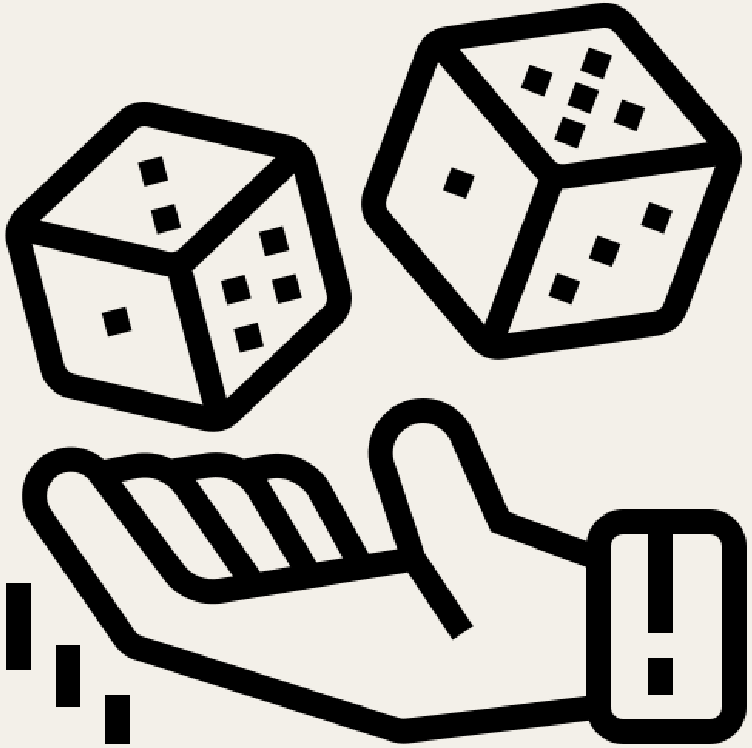
The problem: Trust

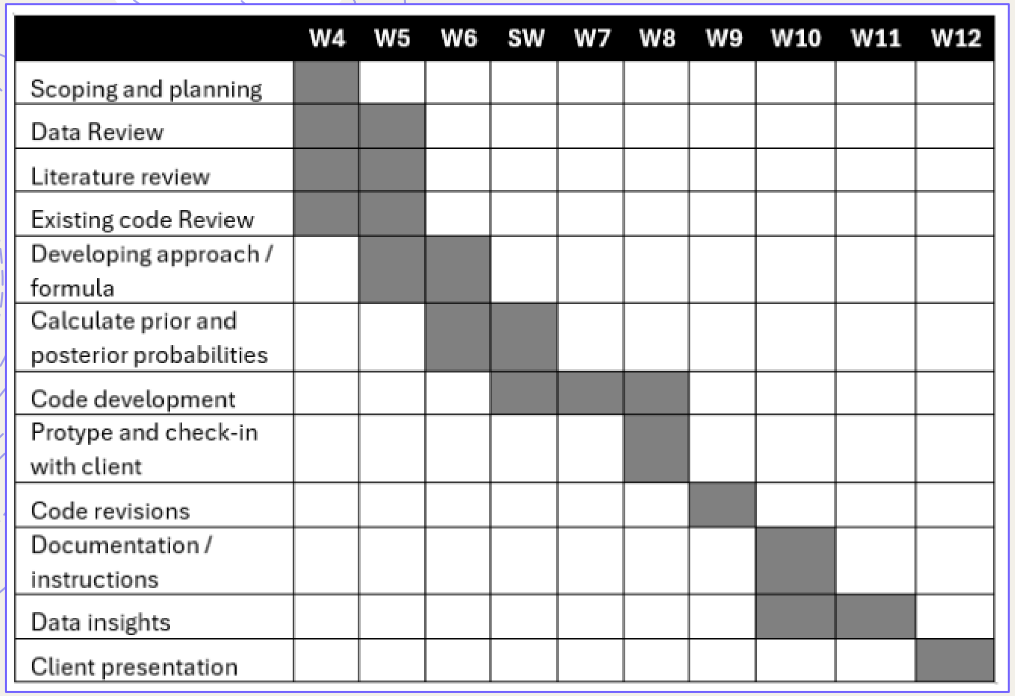
- New / evolving technology
- Sample contamination
- Data processing errors
- Reliability of DNA databases
- Varied rates of species DNA divergence

The data science solution

Bayesian model using MCMC simulations to generate a probability for whether the BLAST DNA match is accurate

- Take in the 1,000,000+ eDNA samples
- Data match with external data sources (GBIS, OBIS, AquaMaps, Fish Tree of Life, Fishbase)
- Treat this as "new" information in a Bayesian context
- Compare with a simplified formula that weights external datapoints to calculate a score





	W4	W5	W6	SW	W7	W8	W9	W10	W11	W12
Scoping and planning	■									
Data Review	■	■								
Literature review	■	■								
Existing code Review	■	■								
Developing approach / formula		■	■							
Calculate prior and posterior probabilities			■	■						
Code development				■	■	■				
Protype and check-in with client						■				
Code revisions							■			
Documentation / instructions								■		
Data insights								■	■	
Client presentation										■

The infallible Gantt chart from our project proposal

The challenges

- Concurrent development hampered by dependencies
- Technical challenges leading to delays
- Competing workload from other projects

What worked

- Task pairs
- Fixed yet flexible meeting schedule
- GIT and Teams
- Support, encouragement and flexibility

Proud to create a functional, adaptable workflow

Achievements

- Enriched dataset with multiple external datapoints
- Bayesian probability model
- Simplified, weighted formula
- Sample level accuracy probability

Widely adaptable

- Finned and cartilaginous fish specific implementation
- Applicable to any other eDNA

Major stages and approach

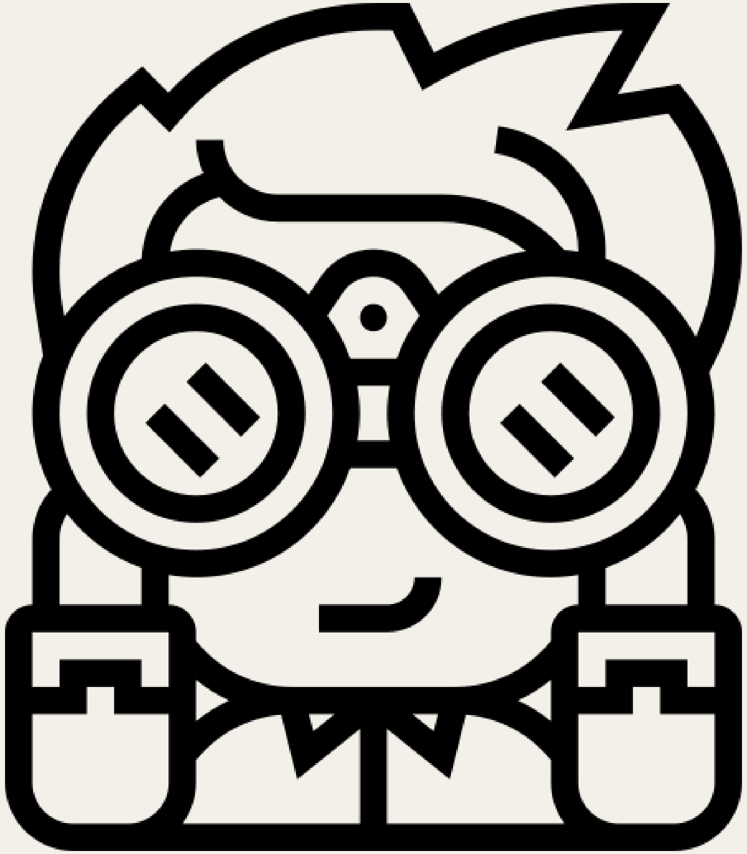
Focus on what matters

- False Positives
- To score each record's reliability by combining DNA evidence with independent location signals at a scale of millions of rows.
- OBIS is a global database of such records

Successful implementation

- Built a scoring pipeline
- Combined signals using naïve bayes
- Safe mode
- EDA visuals and csv outputs

Motivated by finding **data efficiencies** + transparent and **reproduceable results**



If I had more time for improvements

Naming normalisation

ML model (XGBoost)

Data cleaning and preparation

Major stages / components

Focus on what matters

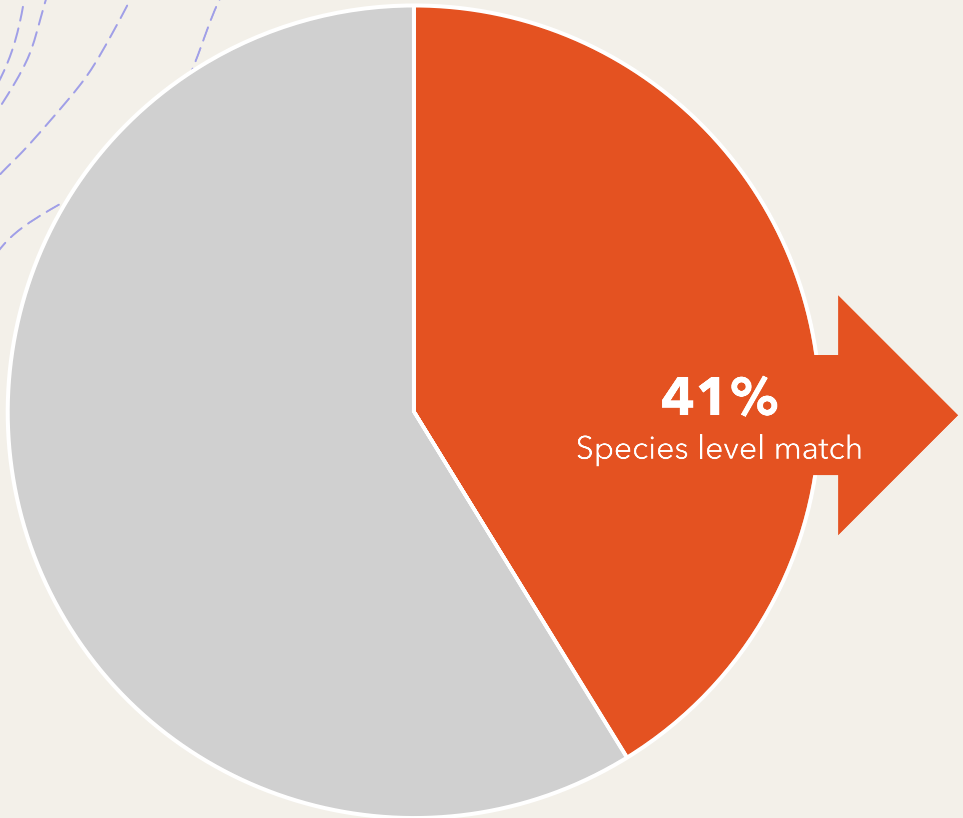
Decrease the size of the data

- Drop irrelevant columns
- Drop rows with excessive NAs
- Remove DNA for non-fish species
- Drop all non-species

Correct mistakes and transform data

- Lattitude / longitude errors
- Normalise diversity rate, counts
- Log scale transform to obtain normal distribution

Processing the 1,067,313 samples in the original dataset



Location Data

GBIF	95%
OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM	95%
AquaMaps	100%

DNA Data

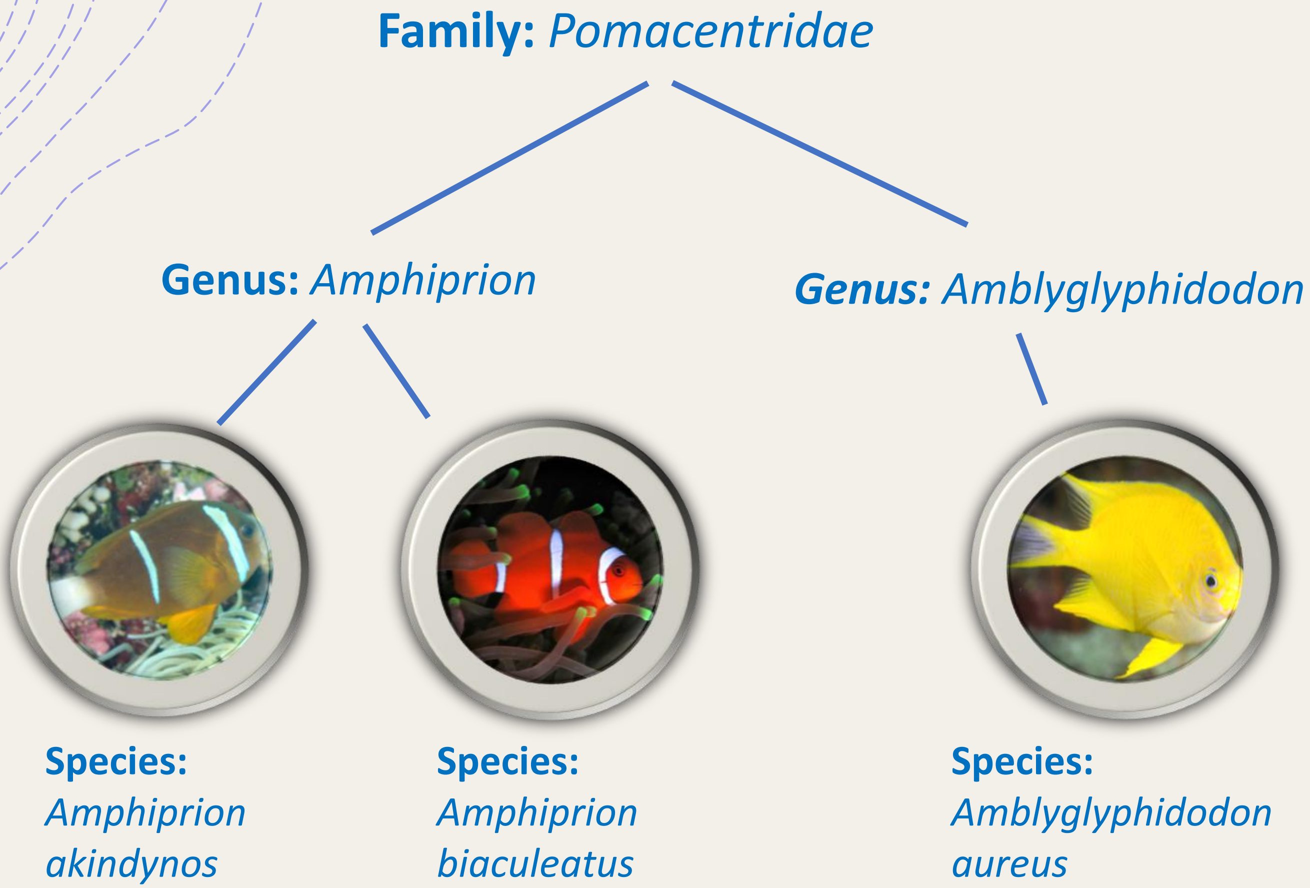
Fish Tree of Life	88%
ENA European Nucleotide Archive	96%

Code Management

Git, GitHub, and Team Workflow

- Delivered team GitHub training
- Created branches for each member
- Created securities protocols for the main branch.
- Combine the repos into a single project
- Rscript to run all the code

Major challenge: Bayesian Formulation



Family: *Pomacentridae*

Genus: *Amphiprion*

Genus: *Amblyglyphidodon*

Species: *Amphiprion akindynos*

Species: *Amphiprion biaculeatus*

Species: *Amblyglyphidodon aureus*

How complex should the model be?

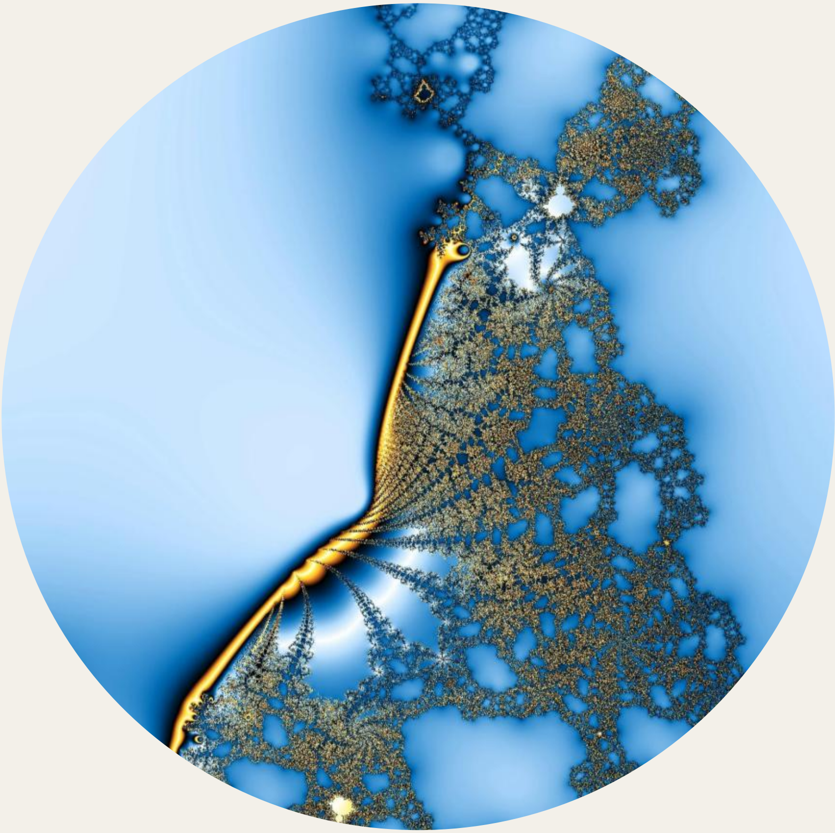
$P(Species \mid Genus \;\cap Location)$?

$P(Genus \mid Family \;\cap Location)$?

$P(Family \mid Domain \;\cap Location)$?

Images from Fishes of Australia

Approaching the challenge at the species level is complex enough!



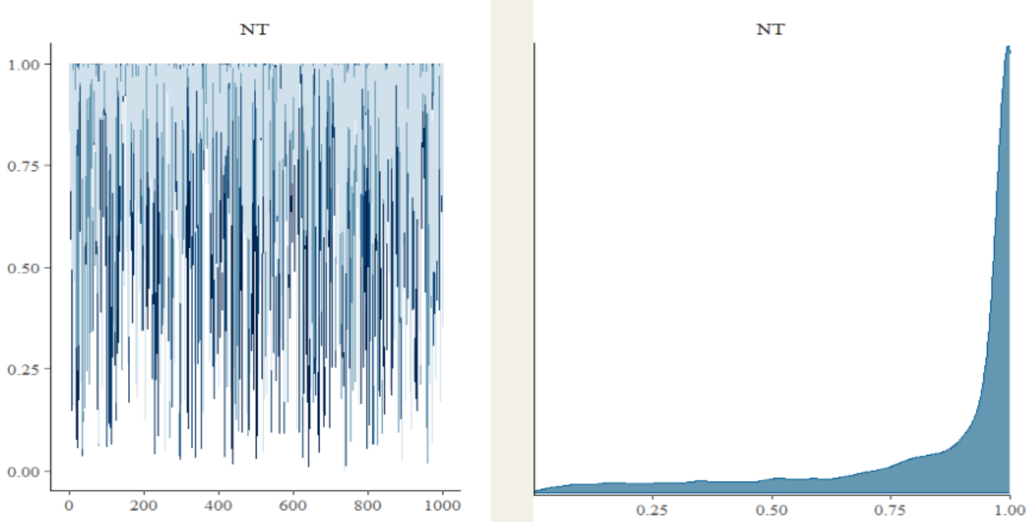
$$P(S\ |ID,RD,GP,DR,AQ,NT,DI) = \frac{P(ID \mid DR,\ \ S)\,P(\ RD \mid S)\,P(GP \mid S)\,P(AQ \mid S\ )\,P(NT \mid S\ )\,P(DI \mid S)\,P(S)}{P(ID \mid DR)\,P(RD)\,P(GP)\,P(AQ)\,P(NT)P(DI)}$$

Estimating the Posterior

To avoid the need to calculate the denominator, rewrite to the form:

$$f(S \mid ID, RD, GP, DR, AQ, NT, DI) \propto f(ID \mid DR,\ S)\, f(RD \mid S)\, f(GP \mid S)\, f(AQ \mid S)\, f(NT \mid S)\, f(DI \mid S)\, f(S)$$

- All probabilities must sum to 1
- Find the expected value of the distribution
- Avoid nasty integrations by using simulations in Stan

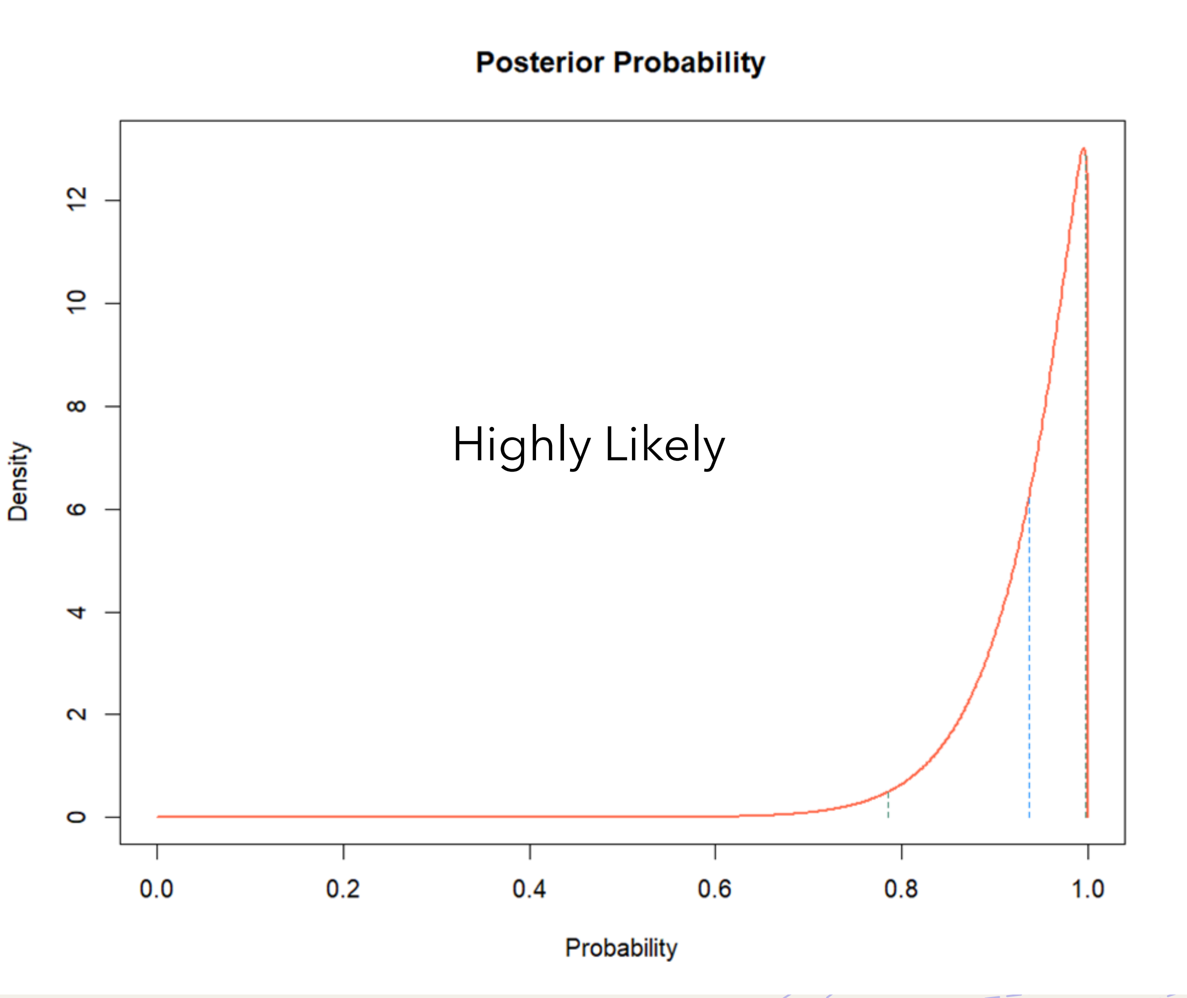


Example of Stan trace and density plots for probability of species given number in target area

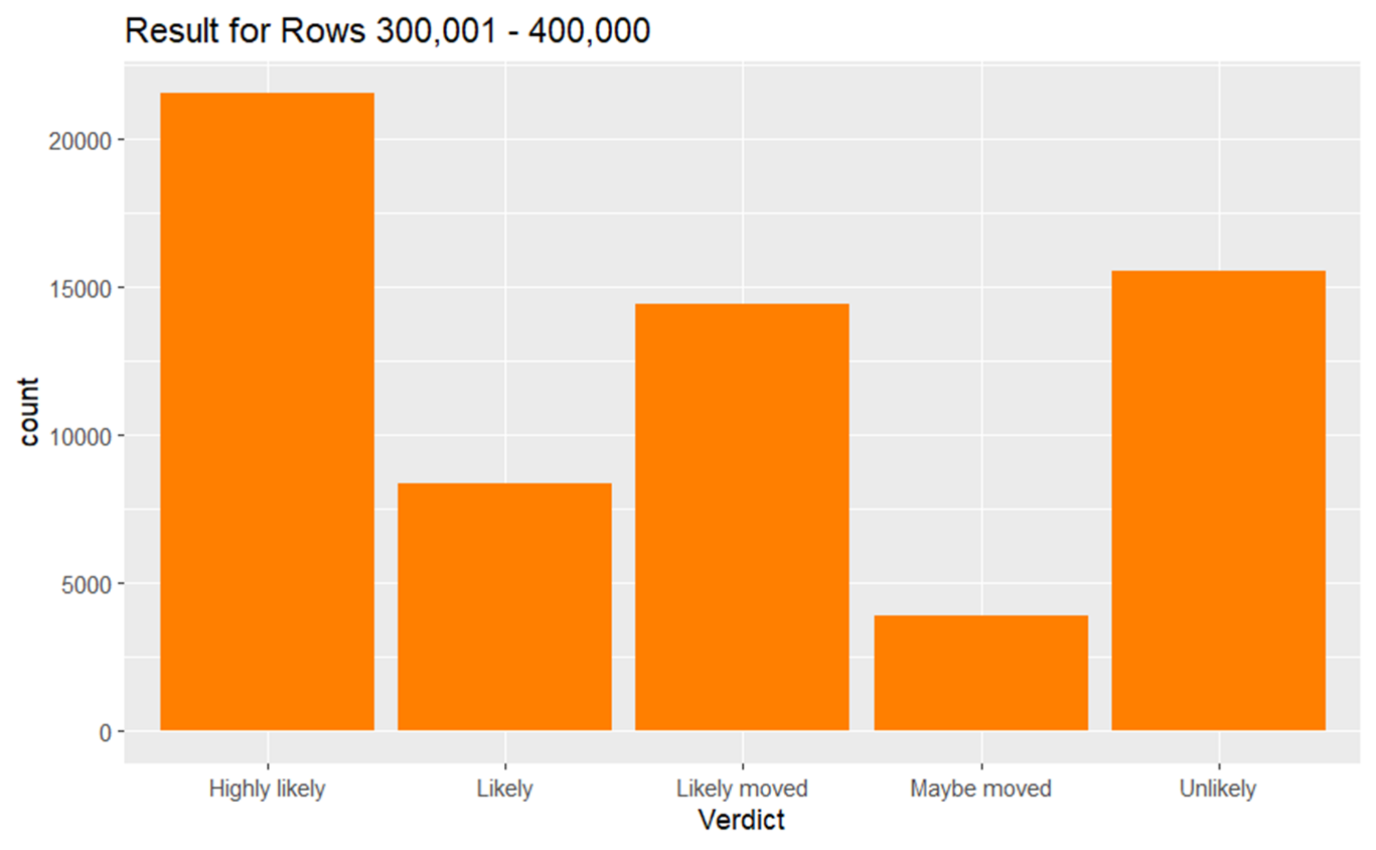
Which Distribution to Use?

Beta distribution:

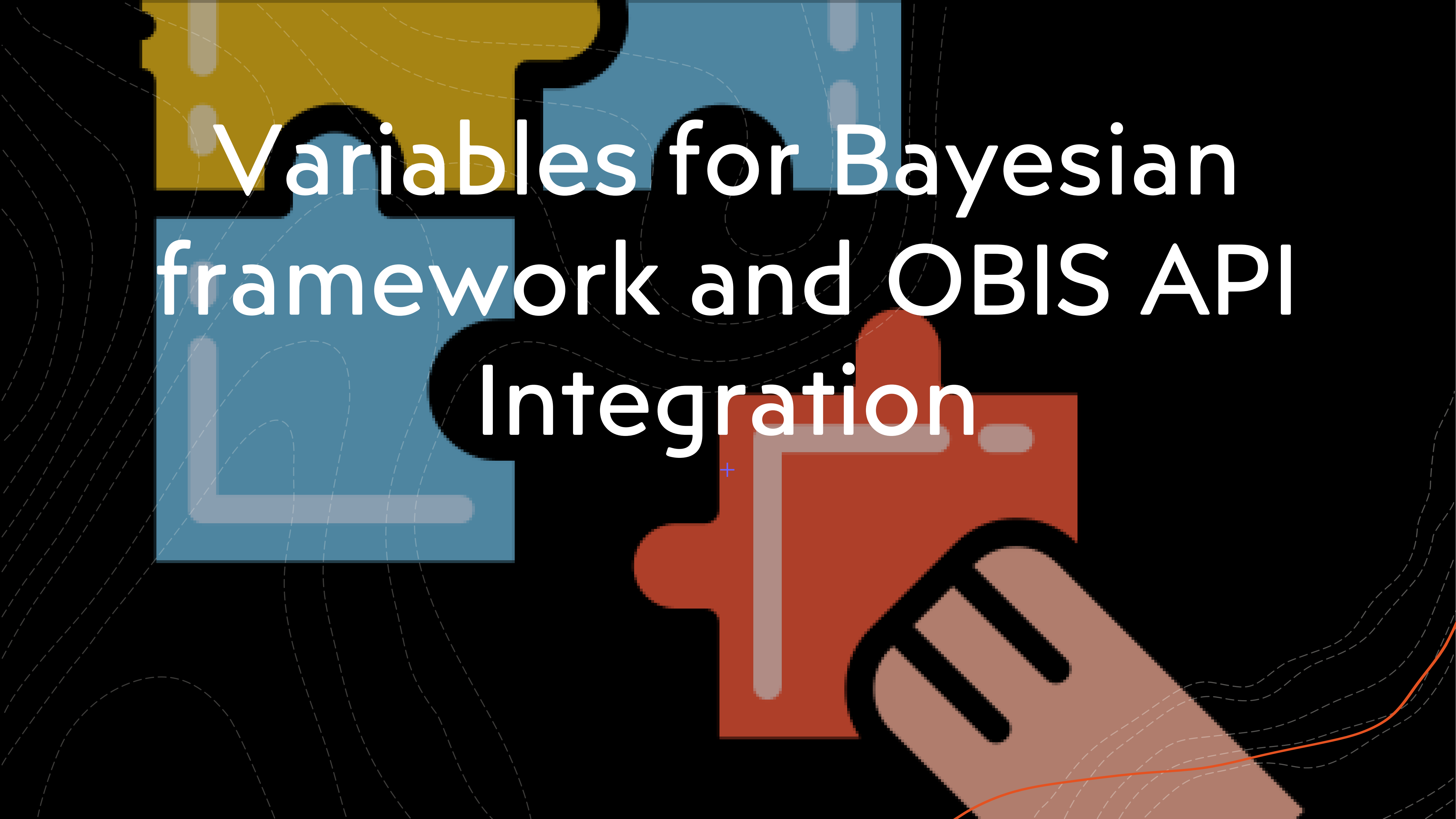
- Has range between 0 and 1
- Total area under the curve is 1
- This is perfect for probabilities!



22% of results so far predicted as accurate

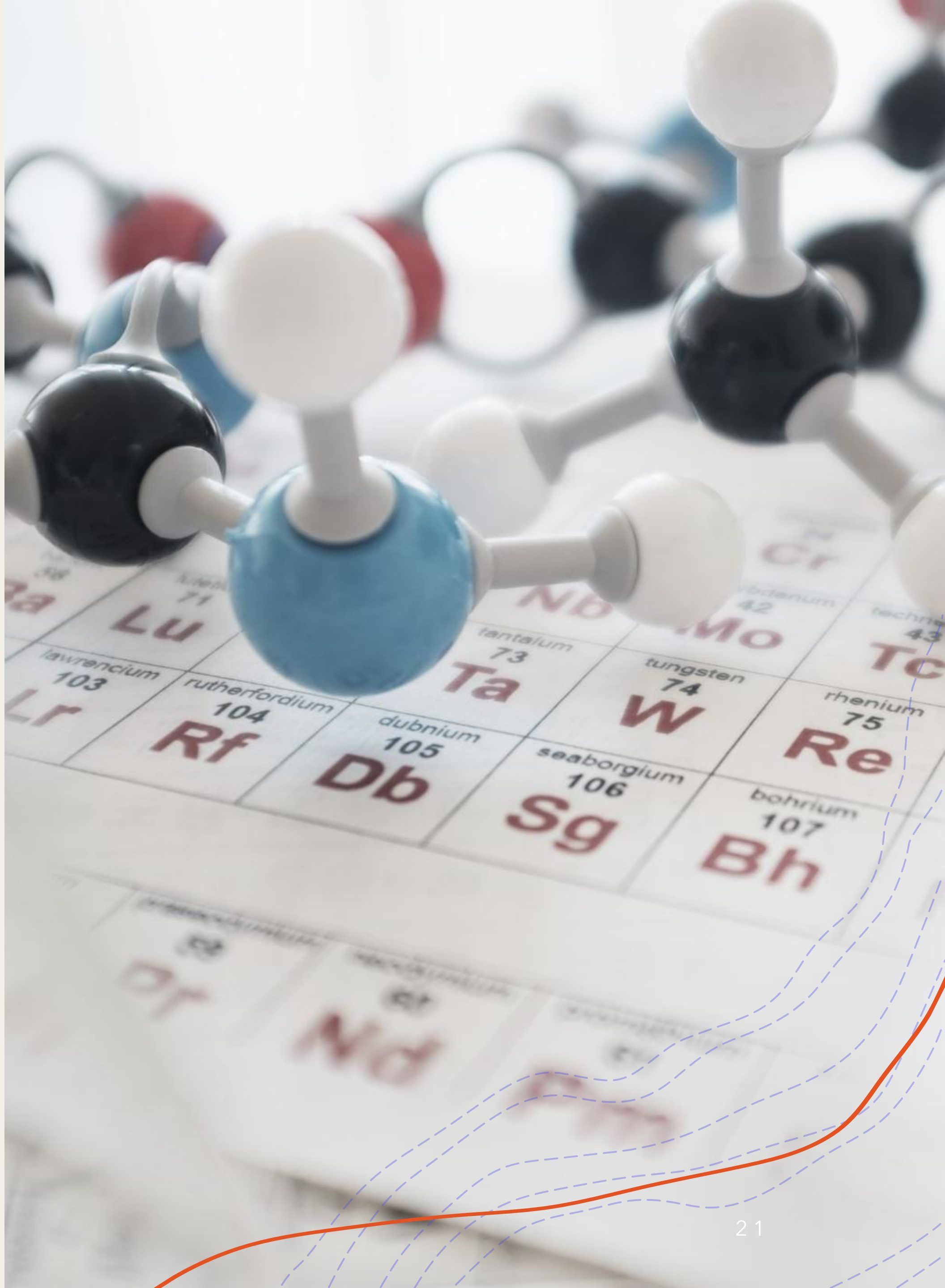


Variables for Bayesian framework and OBIS API Integration



Technical Challenge

- Traditional eDNA methods rely solely on BLAST similarity scores above 97% to confirm species detections.
- Many species are missing from reference databases, which leads to false positives and misidentifications.
- Some DNA markers cannot distinguish between very closely related species.
- A probabilistic model was needed that combines genetic evidence, taxonomic precision, and ecological plausibility.
- The OBIS API was integrated to verify whether detected species are geographically plausible at sampling locations.



Bayesian Framework

Sequence Evidence

- **This** component includes percent identity, read counts, and assay type, which feed the likelihood in the Bayesian model.

Taxonomic Context

- **Species**, genus, family, and Lowest Common Ancestor prevent overconfidence when evidence only supports genus-level classification.

Ecological Plausibility

- **Latitude** and longitude coordinates are used to query external databases like OBIS, FishBase, and AquaMaps to create geographic priors.

Data Quality

- **Filters** were applied to remove weak signals or low read counts that could represent sequencing noise..

OBIS API Integration

- Automated queries were performed for each detection, sending both species names and sample coordinates to OBIS.
- A bounding box with approximately 100 kilometers buffer was created around Australia to capture coastal and offshore records.
- Geographic priors were increased when species occurrence records were found near the sampling location.

Motivation

I wanted to improve the reliability of environmental DNA identification and reduce the occurrence of false positives.

Traditional workflows rely on high BLAST matches without verifying ecological plausibility, which needed to be addressed.

I was motivated to learn and apply Bayesian reasoning and probabilistic thinking to real biological problems.

This project provided an opportunity to merge ecology, genetics, and computation into one coherent analytical framework.

Outcome

- I successfully developed a working OBIS API integration that performs real-time species presence checks for each detection.
- A complete variable framework was designed that unites genetic match quality, taxonomic precision, ecological plausibility, and data reliability.
- The Bayesian scoring approach significantly reduced false positives compared to traditional threshold-based methods.
- Results became more interpretable through probability-based confidence scores rather than binary yes-or-no classifications.

Future Improvements

+ I plan to develop a Shiny dashboard that summarizes detection results at the voyage level for interactive analysis.
+ The dashboard will highlight the most common species detected in each voyage to identify biodiversity patterns.
+ It will flag the rarest taxa—species with very low global OBIS occurrence but strong DNA evidence—to reveal potentially significant finds.
+ The system will display the nearest known OBIS record for every detection to visualize how far each sighting is from known distributions.
+ This would transform the system from a statistical model into a practical decision-support tool for marine ecologists.

From Database to Bayesian Priors: The OBIS Journey

Project challenges

1. Understand the Pipeline
How does eDNA flow from FASTQ → OceanOmics → LCA → Results?

2. Find the Right Database
Which database gives us occurrence frequency (not just yes/no)?

Bayesian models need **good priors**—occurrence data, not assumptions!

Motivation

Learning & Skill Development

- Opportunity to know about marine biodiversity informatics
- Develop expertise with new tools and databases Build foundational knowledge

Personal growth

- Push beyond my comfort zone into unfamiliar territory
- Bridge the gap between theory (Bayesian stats) and practice

Approach

Research Foundation

- Kaehler et al. (2019): 25% → 14% error with proper priors
- This proved occurrence data was the key

Database Comparison Implementation

- Created OBIS queries for **species occurrence counts**
- Built functions to find **nearest distance** to known occurrences
- Developed **species name cleaning** for accurate matching
- Documented learnings for team handoff for further implementations

Two wins and a Future Improvement

OBIS

- Learned a new database
- Successful API integration

Theory → Practice

- Connected Bayesian statistics to database selection
- Found the data that makes better priors possible

Deeper on sources

Vision

- Multi-database validation

Additional Sources to Explore

- Atlas of Living Australia (ALA) - more localized WA data
- Museum specimen collections - verified records
- Regional marine databases - fill coverage gaps

Technical responsibilities

API Implementation

Location Processing

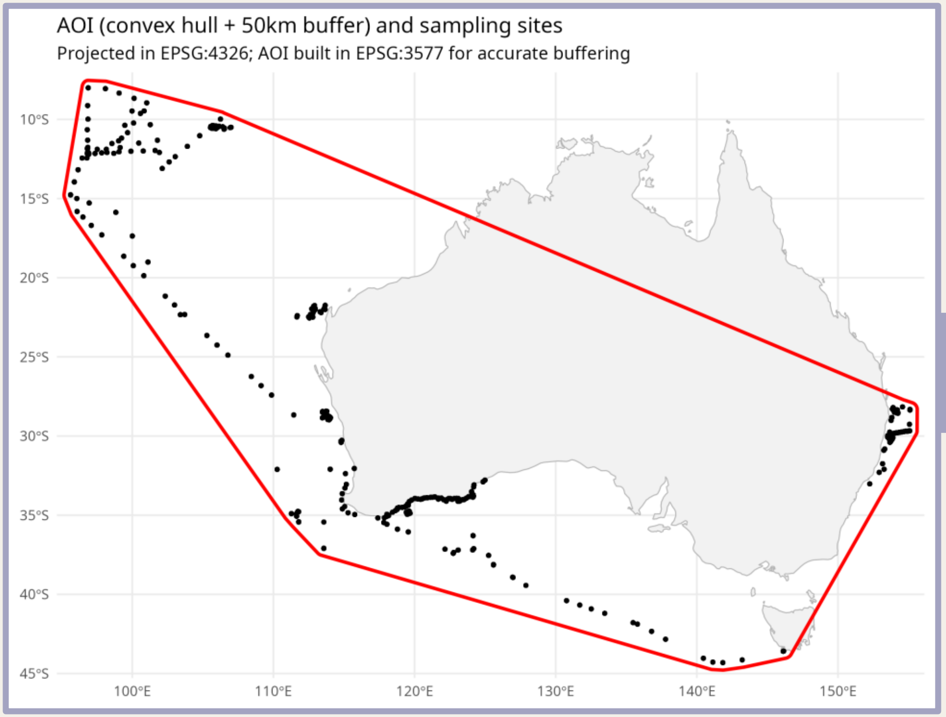
Local Data Cache

Driven by the real-world application and contributing to help others who are protecting the environment

GBIF API was not as simple as first thought...

- Name variations
- Millions of observations
- Query rate limits

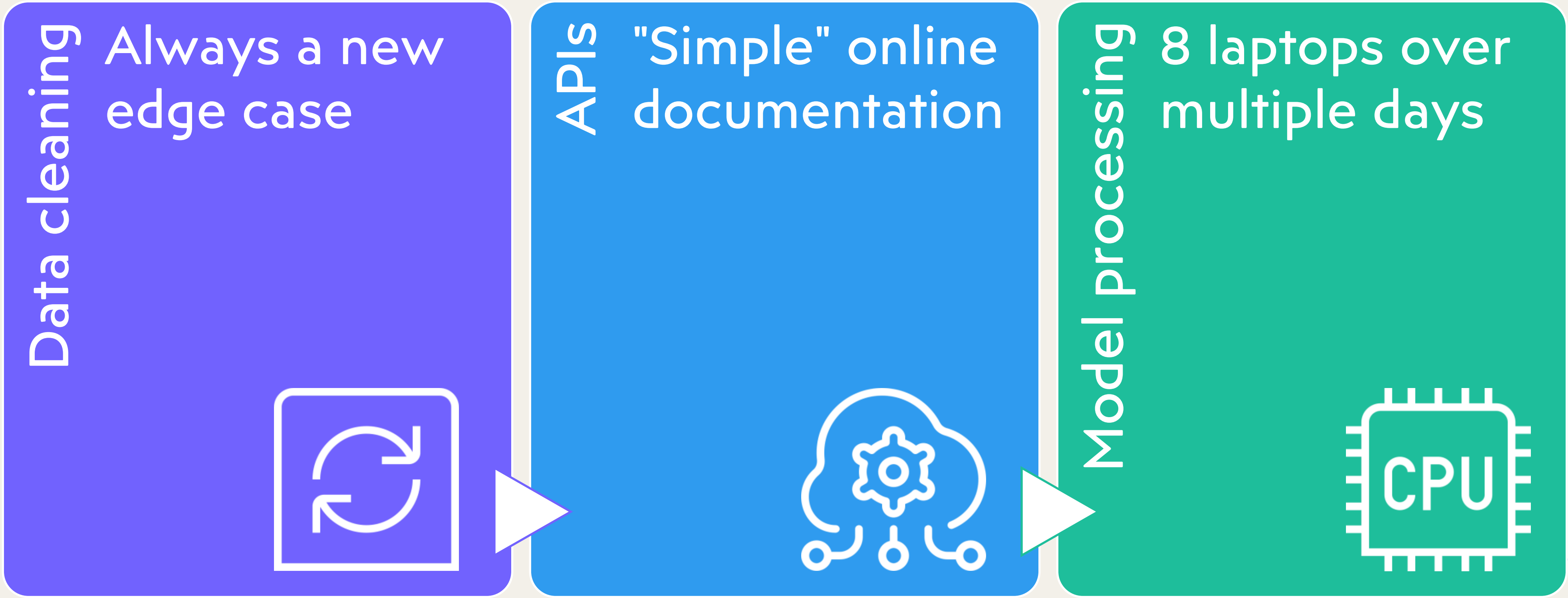
GBIF Pipeline: Extract sample locations >> Project an area of interest >> build local cache of species observations + DNA



AOI (convex hull + 50km buffer) and sampling sites

Projected in EPSG:4326; AOI built in EPSG:3577 for accurate buffering

Proud of the team and their attitude in response to each new challenge



Data cleaning

Always a new edge case

APIs

"Simple" online documentation

Model processing

8 laptops over multiple days

Genus level match processor is nearly complete…

- For BLAST DNA matches where a genus was found but not a specific species
- Explodes 440k > 1.6m rows
- Apply model and select species with highest probability

Thank You