# Age of Abalone Prediction

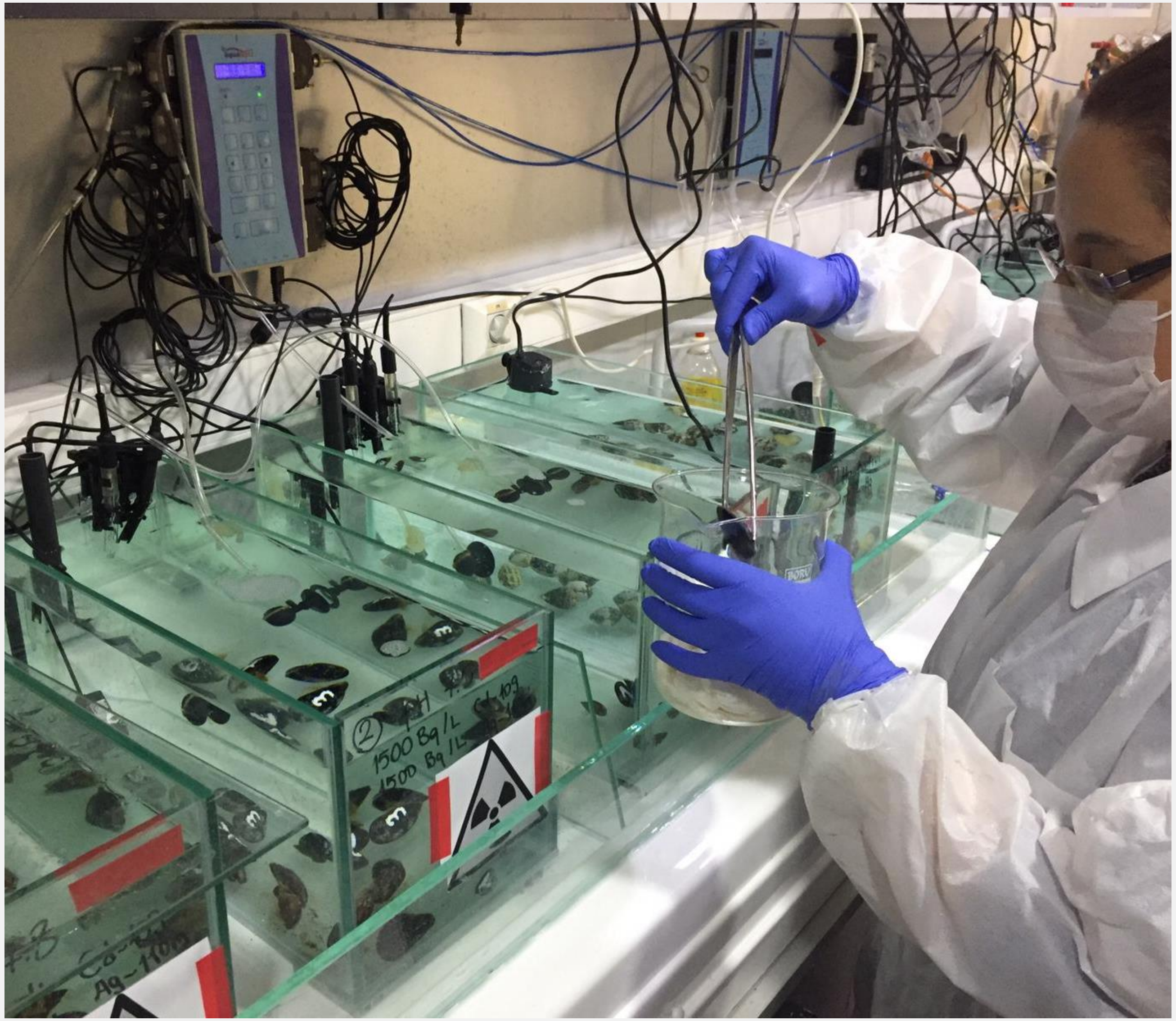

Department of Primary Industry and Fisheries of Tasmania

Amanda Gregory - Claire Xiao - Jason Qi - Yves Esslinger

# What is our goal?

Create a model with improved accuracy that predicts the age of abalone

- Simplify and shorten the long, tedious process to discover the age of abalone.
- Information proves valuable to researchers and consumers



# Approach

Data → Initial Analysis → Evaluating Previous Work → Comparison of Different Models → Conclusions

# Abalone Dataset

| Sex | Length | Diameter | Height | WholeWeight | ShuckedWeight | VisceraWeight | ShellWeight | Rings |
|---|---|---|---|---|---|---|---|---|
| M | 0.455 | 0.365 | 0.095 | 0.514 | 0.2245 | 0.101 | 0.15 | 15 |
| M | 0.35 | 0.265 | 0.09 | 0.2255 | 0.0995 | 0.0485 | 0.07 | 7 |
| F | 0.53 | 0.42 | 0.135 | 0.677 | 0.2565 | 0.1415 | 0.21 | 9 |
| M | 0.44 | 0.365 | 0.125 | 0.516 | 0.2155 | 0.114 | 0.155 | 10 |
| I | 0.33 | 0.255 | 0.08 | 0.205 | 0.0895 | 0.0395 | 0.055 | 7 |


| Male | Female | Intersex | Length | Diameter | Height | WholeWeight | ShuckedWeight | VisceraWeight | ShellWeight | Rings |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0.455 | 0.365 | 0.095 | 0.514 | 0.2245 | 0.101 | 0.15 | 15 |
| 1 | 0 | 0 | 0.35 | 0.265 | 0.09 | 0.2255 | 0.0995 | 0.0485 | 0.07 | 7 |
| 0 | 1 | 0 | 0.53 | 0.42 | 0.135 | 0.677 | 0.2565 | 0.1415 | 0.21 | 9 |
| 1 | 0 | 0 | 0.44 | 0.365 | 0.125 | 0.516 | 0.2155 | 0.114 | 0.155 | 10 |
| 0 | 0 | 1 | 0.33 | 0.255 | 0.08 | 0.205 | 0.0895 | 0.0395 | 0.055 | 7 |

# Data

## Sex

- Description: M (Male), F (Female), I (Immature)
  - Used binary coding to define Sex (1, 0)
- Data Type: Nominal

## Length

- Description: Longest shell measurement in mm
- Data Type: Continuous
- Statistics: Min-0.075, Max- 0.815, Mean- 0.524, SD- 0.120

## Diameter

- Description: Longest shell measurement perpendicular to length in mm
- Data Type: Continuous
- Statistics: Min-0.055, Max- 0.650, Mean- 0.408, SD- 0.099

## Height

- Description: Measurement of height with meat in shell in mm
- Data Type: Continuous
- Statistics: Min-0.000, Max- 1.130 , Mean- 0.140, SD- 0.490

## Whole Weight

- Description: Weight of entire abalone in grams
- Data Type: Continuous
- Statistics: Min-0.002, Max- 2.826, Mean- 0.829, SD- 0.490


Measurement Tool for Abalone

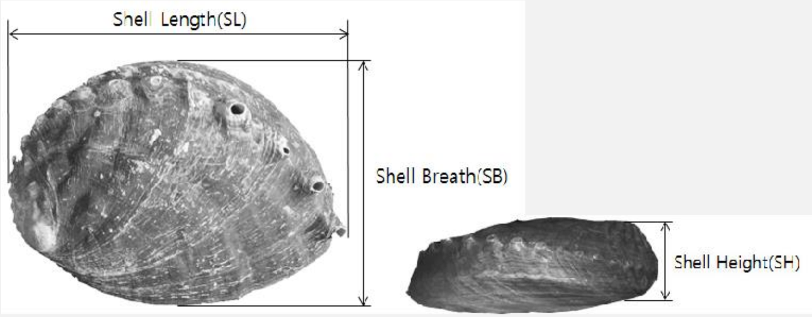



Diagram Displaying Length (SL), Diameter (SB) and Height (SH) of an Abalone

# Data

## Whole Weight

- Description: Weight of abalone meat in grams
- Data Type: Continuous
- Statistics: Min-0.001, Max- 1.488, Mean- 0.359, SD- 0.222

## Viscera Weight

- Description: Gut weight of abalone (after bleeding) in grams
- Data Type: Continuous
- Statistics: Min-0.001, Max- 0.760, Mean- 0.181, SD- 0.110

## Shell Weight

- Description: Weight of abalone shell after being dried in grams
- Data Type: Continuous
- Statistics: Min-0.002, Max- 1.005, Mean- 0.239, SD- 0.139

## Rings

- Description: Number of rings in an abalone shell. The age is determined to be the number of rings+1.5.
- Data Type: Integer
- Statistics: Min-1, Max- 29, Mean- 9.934, SD- 3.224

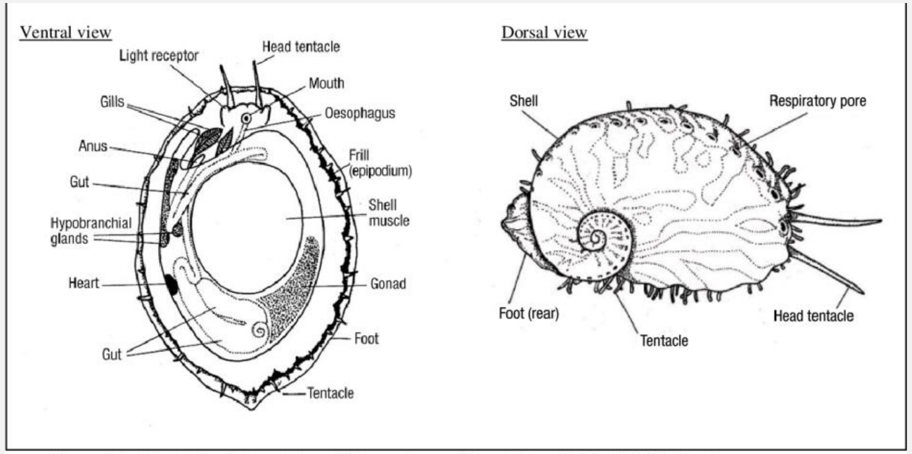
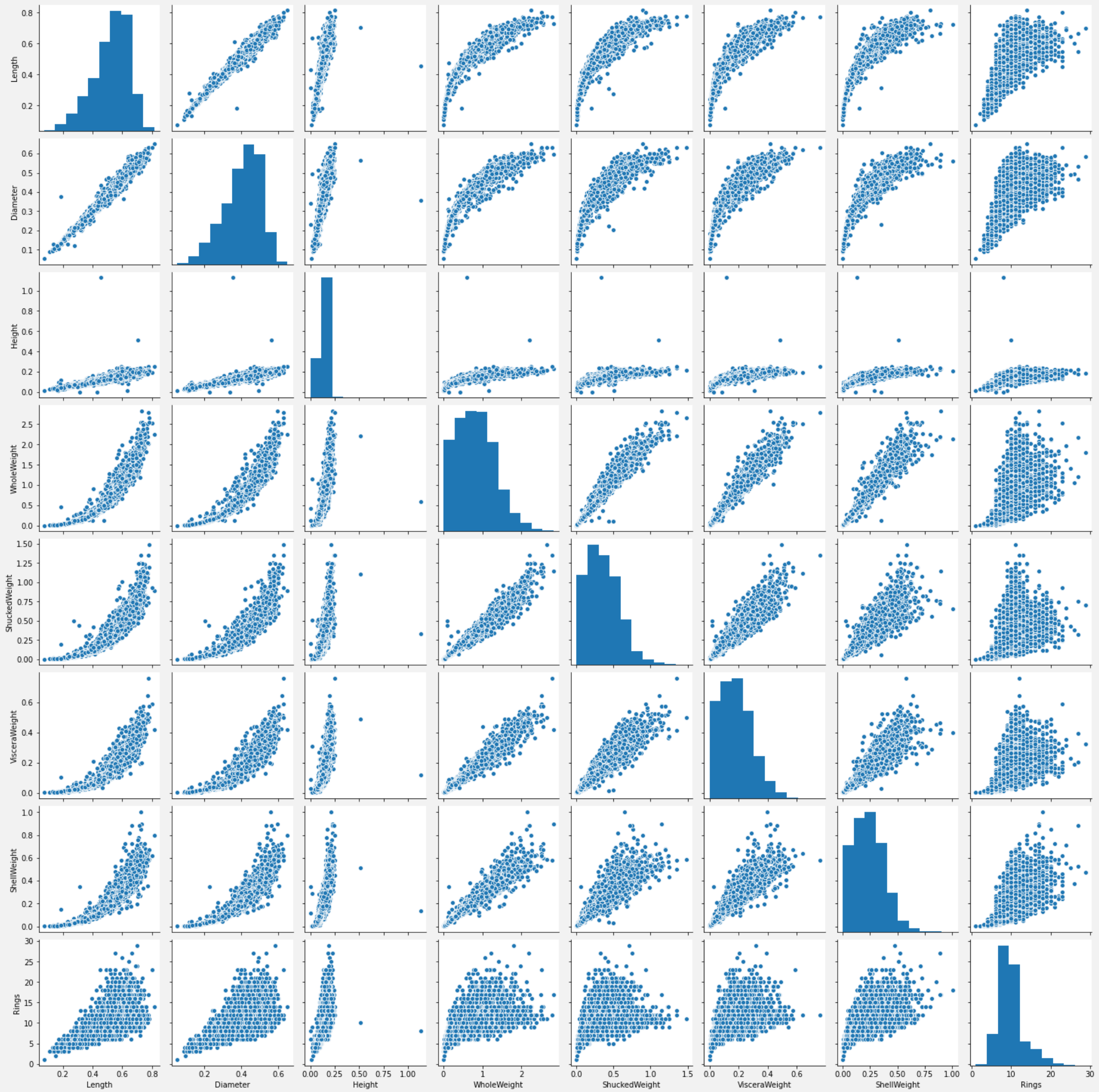


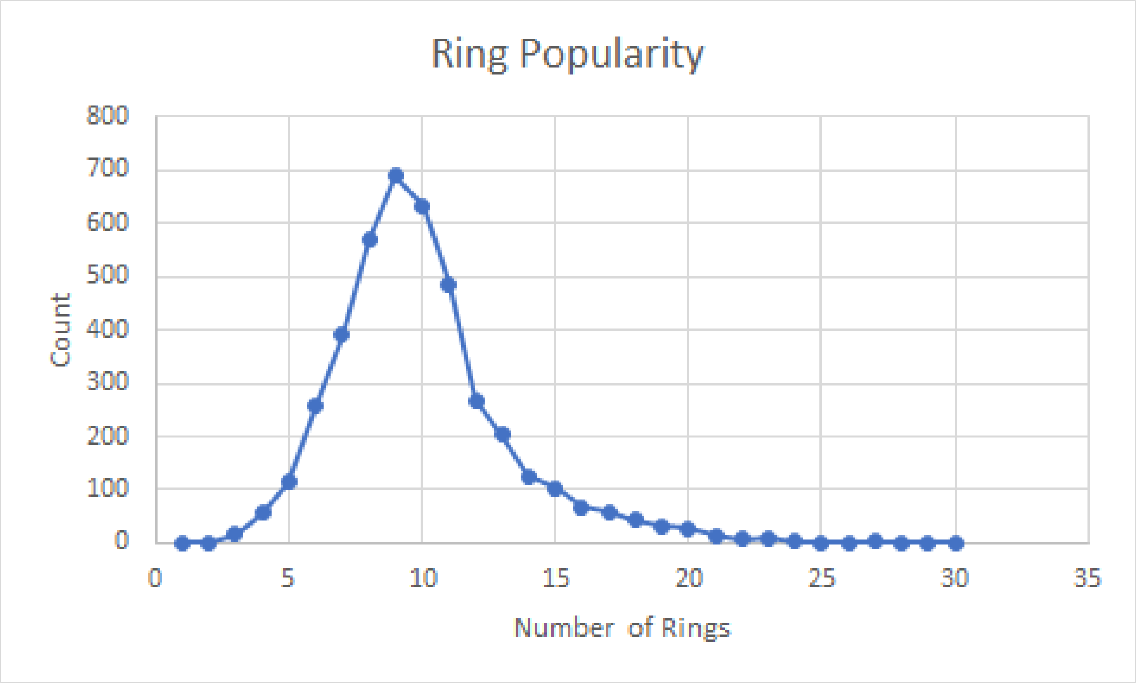

**Figure 1.1. Ventral and dorsal view of the anatomy of the abalone (Fallu, 1994 [online]).**

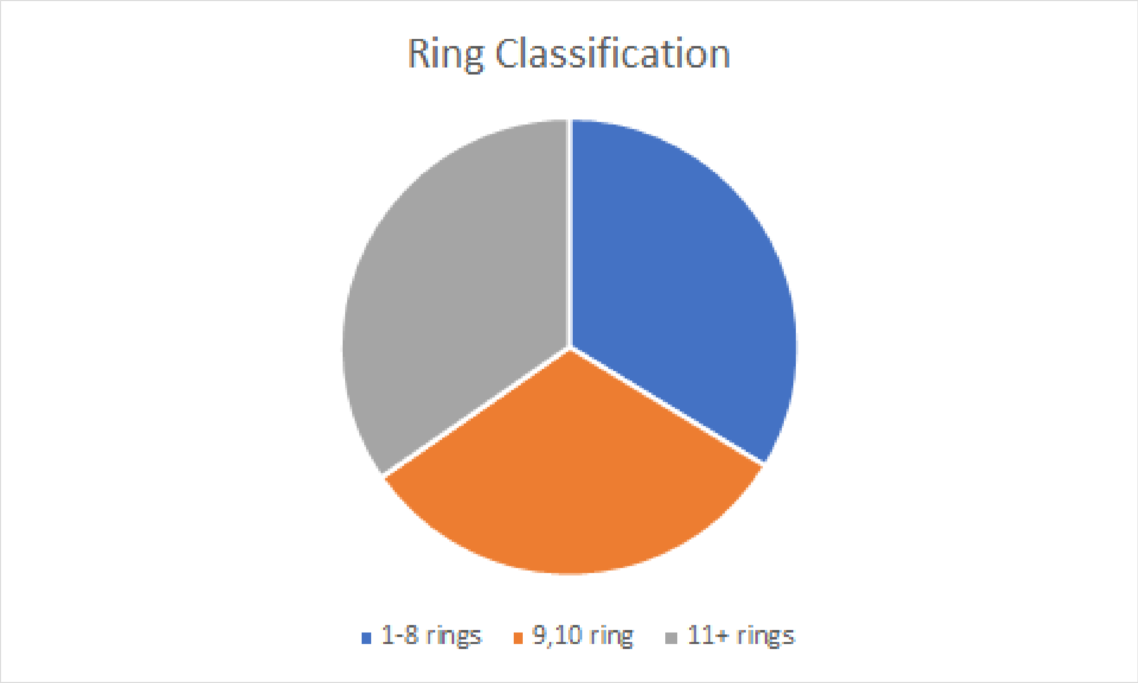
# Pair Plot

- Gender omitted
- Diagonal displays histograms of data points in data set
- Other plots display correlation between two variables
  - All interaction between variables display positive correlation
  - Number of rings is positively correlated with each variable, indicative of

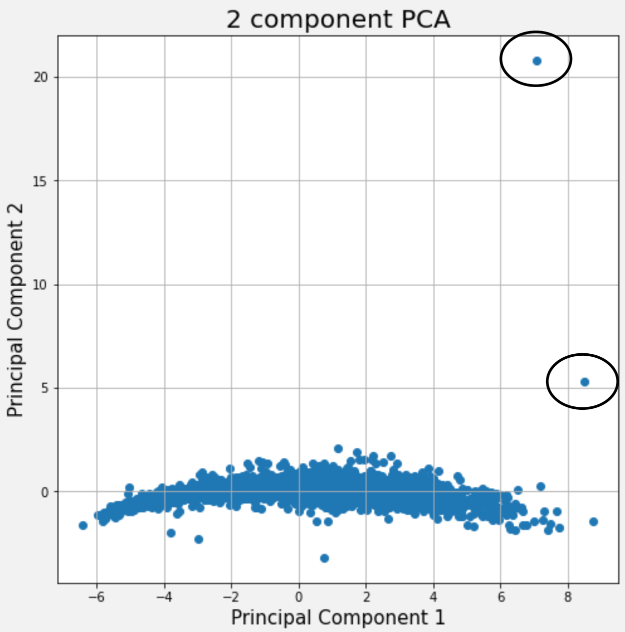


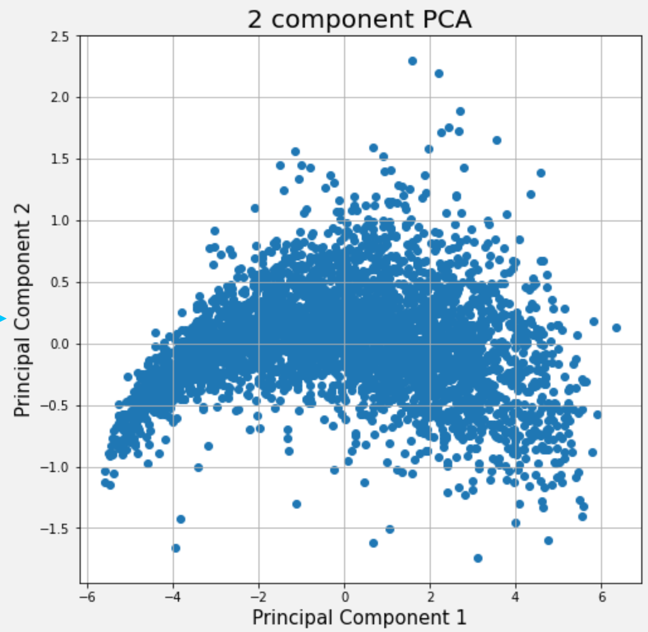
# Data Classification

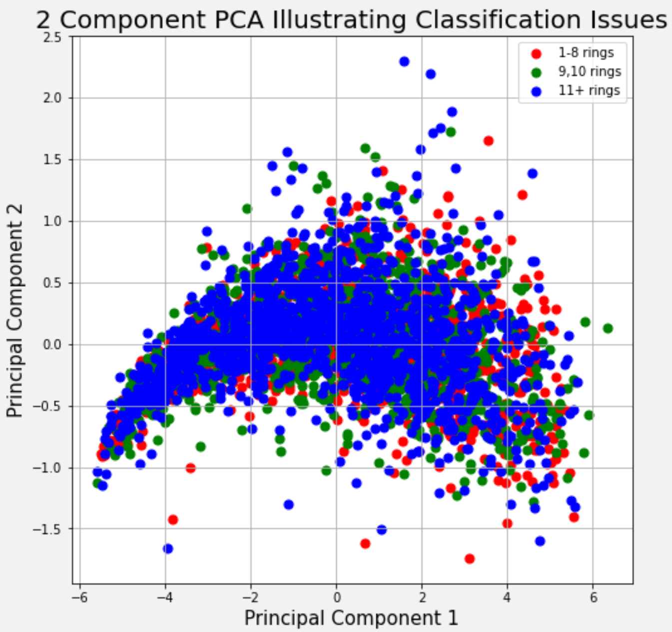


# Closer Look at Dataset Using PCA



# Previous Work on Abalone Data

**Approach**

- Classification

**Idea**

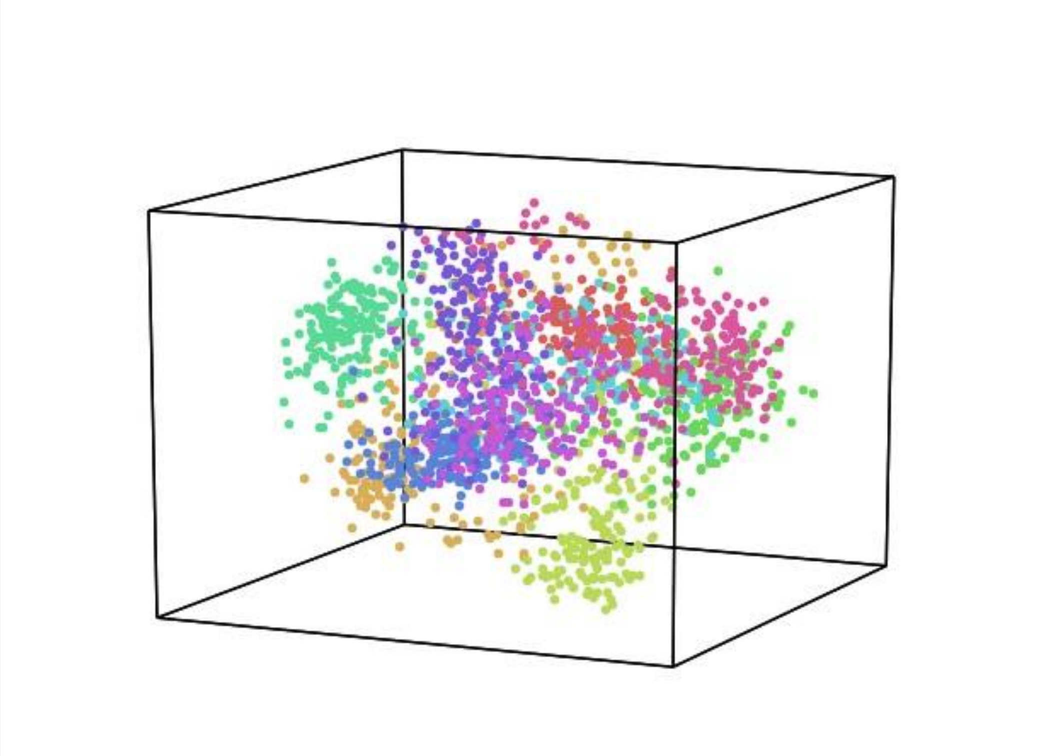
- Extending and Benchmarking Cascade-Correlation

**Author(s)**

- Sam Waugh

**Test Set Performance**

- 3133 training, 1044 testing
- 24.86% Cascade-Correlation (no hidden nodes)
- 26.25% Cascade-Correlation (5 hidden nodes)
- 21.5% C4.5
- 0.0% Linear Discriminate Analysis
- 3.57% k=5 Nearest Neighbor
- Data set samples are highly overlapped.



*Visualization not representative of data*

# Previous Work on Abalone Data

## Approach

- Grouped Classification
  - Group 1: ring classes 1-8
  - Group 2: ring classes 9 and 10
  - Group 3: ring classes 11+

## Idea

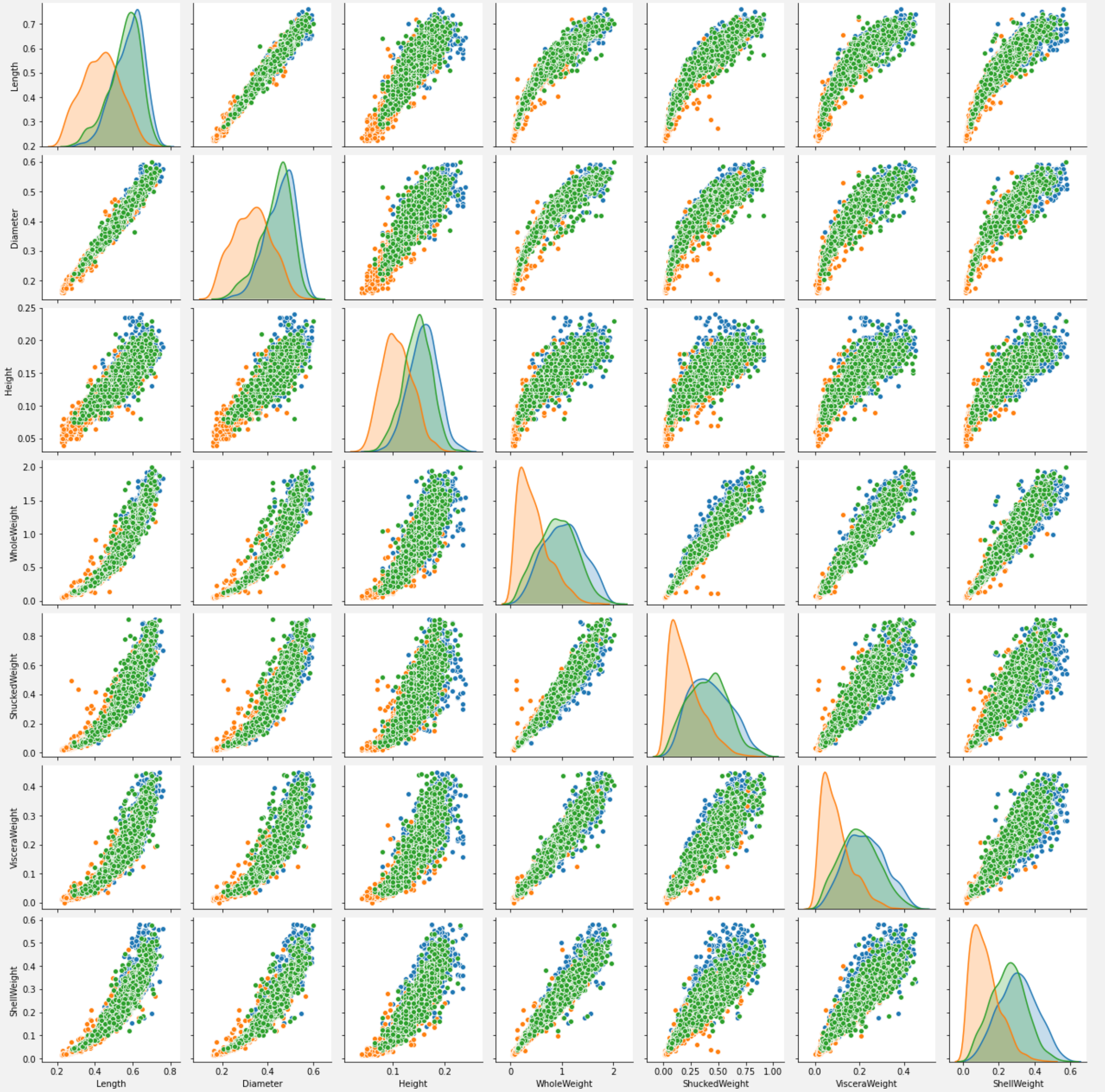
- Comparing Dystal and Backpropagation

## Author(s)

- David Clark, Zoltan Schreter, Anthony Adams

## Test Set Performance

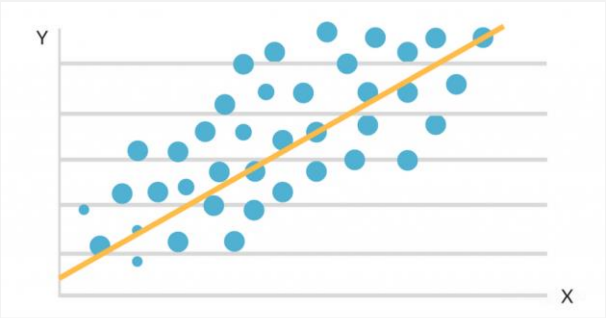
- 3133 training, 1044 testing
- 64%    Backprop
- 55%    Dystal
- 61.40% Cascade-Correlation (no hidden nodes)
- 65.61% Cascade-Correlation (5 hidden nodes)
- 59.2%  C4.5
- 32.57% Linear Discriminant Analysis
- 62.46% k=5 Nearest Neighbour

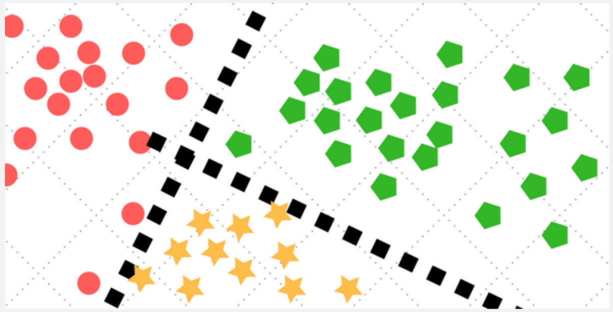



*Orange = 1-8 rings*
*Green = 9,10 rings*
*Blue = 11+ rings*

# General Approach

| | |
|---|---|
| Regression | • Treat the number of rings as a continuous variable |
| Classification by Number of Rings | • Treat the number of rings as a discrete variable<br>• Each number is defined as a specific class |
| Classification by Grouped Number of Rings | • Group the number of rings into three categories<br>• Predict the group assignments |



# Models Used

## Regression

- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- Multi-layer Perceptron regressor
- Support Vector Regressor

## Classification

- Random Forest Classifier
- Multi-layer Perceptron Classifier
- Support Vector Classifier

# Fine-Tuning

- Random Search over possible parameters
- K-fold Cross Validation, k = 5 | number of iterations = 200

```
from sklearn.model_selection import RandomizedSearchCV
kernel = ['linear', 'poly', 'rbf', 'sigmoid']
gamma = ['auto', 'scale']
C = [1, 4, 7, 10, 20, 30]
degree = [2, 3, 4, 5]
shrinking = [True, False]
probability = [True, False]
decision_function_shape = ['ovo', 'ovr']
max_iter = [int(x) for x in np.linspace(5000, 10000, num = 5)]
# Create the random grid
random_grid = {'kernel': kernel,
               'gamma': gamma,
               'decision_function_shape': decision_function_shape,
               'probability': probability,
               'degree': degree,
               'C': C,
               'shrinking': shrinking,
               'max_iter': max_iter}
pprint(random_grid)
```

# Results

| | Regression (metrics = R^2) | Classification (metrics = accuracy) | Grouped Classification (metrics = accuracy) |
|---|---|---|---|
| **Random Forest** | 0.567 vs 0.502 | 0.278 vs 0.244 | 0.641 vs 0.482 |
| **Linear - Lasso** | 0.519 vs 0.301 | - | - |
| **Linear - Ridge** | 0.518 vs 0.514 | - | - |
| **MLP** | 0.575 vs 0.560 | 0.282 vs 0.265 | 0.662 vs 0.655 |
| **SVC** | 0.551 vs 0.488 | 0.266 vs 0.256 | 0.626 vs 0.617 |
| **Best Previous** | - | 0.263 | 0.656 |

# Conclusion

Exhibited multiple models that performed better than previous attempts

MLP with proper fine-tuning performs better than other models

Fine-tuning improves the models' performance

- grid search may improve the result further

Linear regression is not performing as well as other models

- possibly because certain assumptions for linear regression is not satisfied by the data set


Thank You