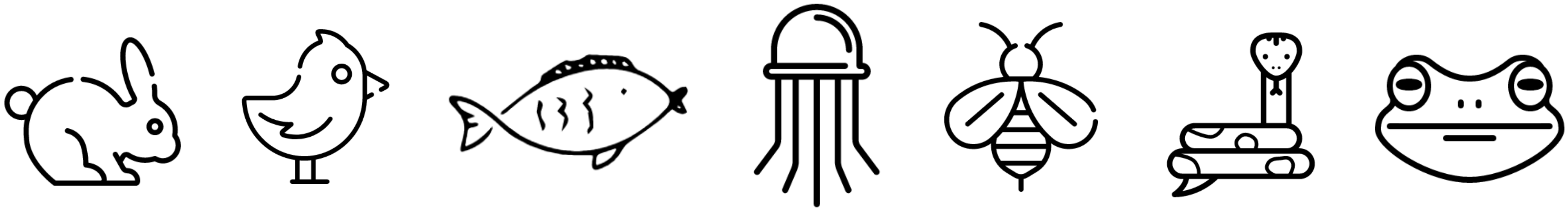
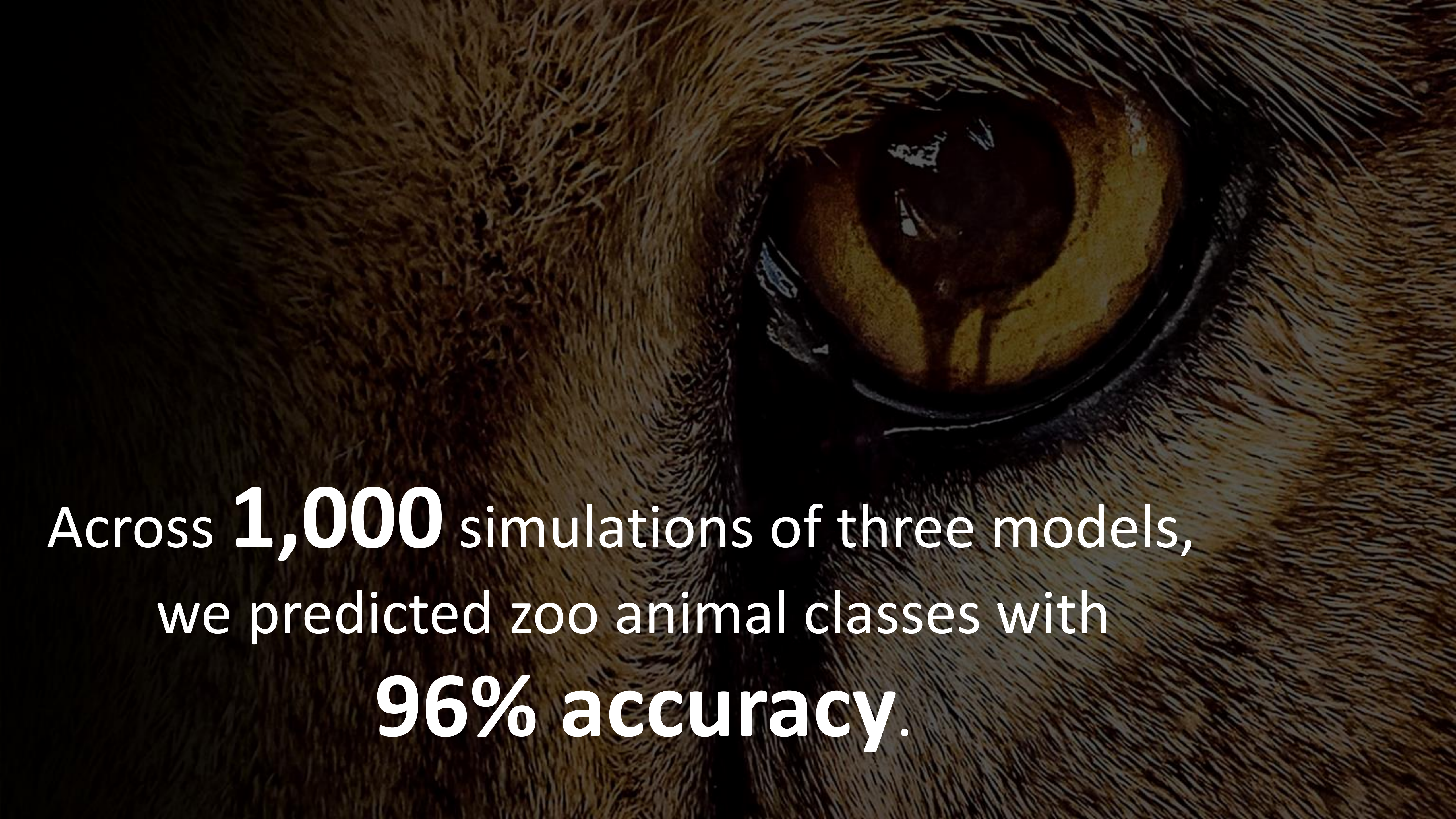
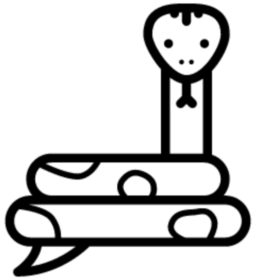
Get This Sea Snake Out of My Zoo!

Sam Ballerini
Clarity Insights Case Study 2018





Across **1,000** simulations of three models, we predicted zoo animal classes with **96% accuracy**.

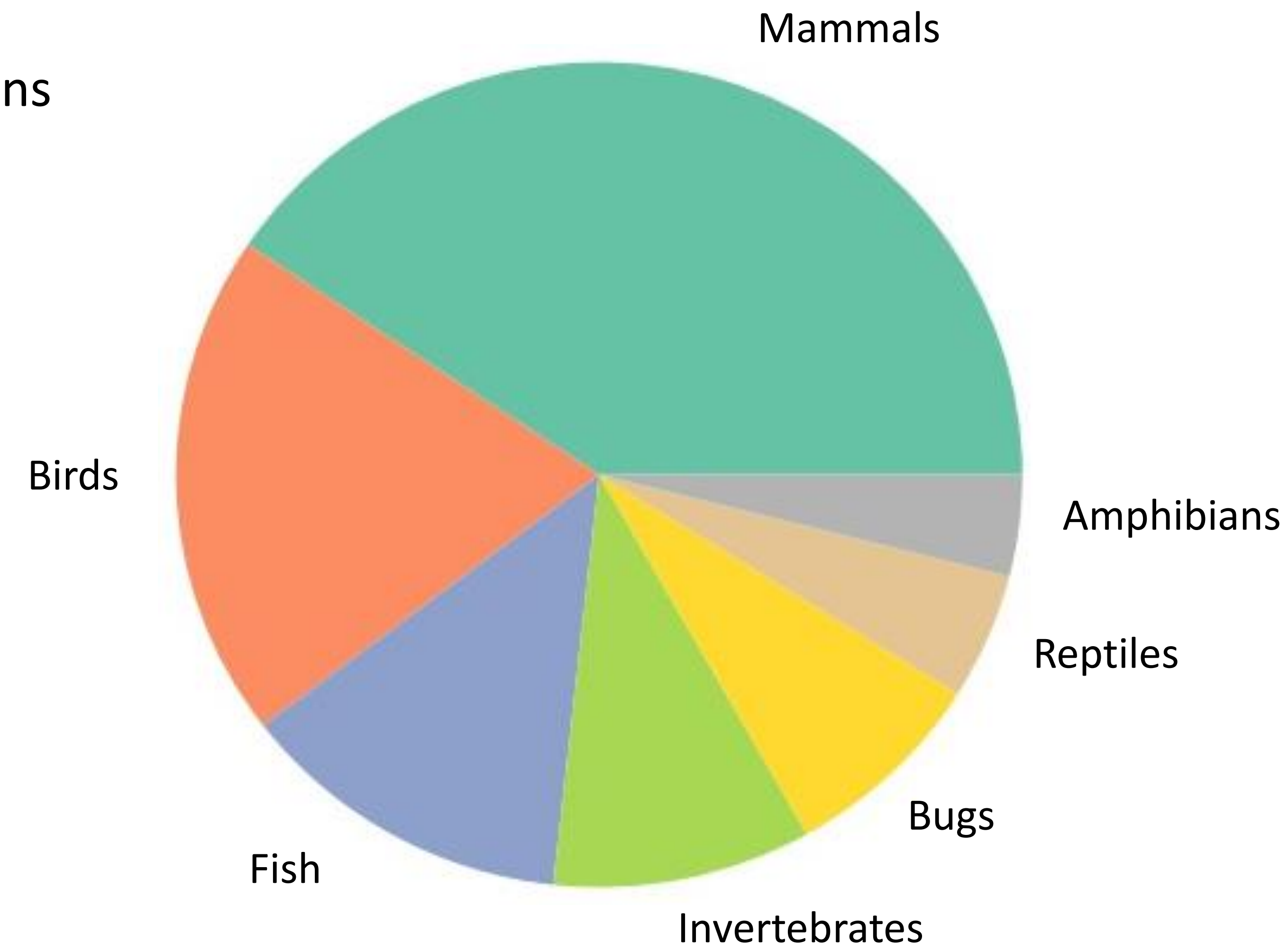


Agenda

1. Problem Overview
2. Data Exploration
3. Model Development
4. Model Evaluation

Class Breakdown

101 observations



Feature Space

Initial Features

- 15 binary variables
- 1 categorical variable with 6 levels (dummy coded)

Feature Space

Initial Features

- 15 binary variables
- 1 categorical variable with 6 levels (dummy coded)

Transformed Features

- 20 binary variables

Feature Space

Initial Features

- 15 binary variables
- 1 categorical variable with 6 levels (dummy coded)

Transformed Features

- 20 binary variables

Variable 1	Variable 2	Pearson Corr.	Counterexample
Milk	Hair	0.88	Dolphin
Feathers	2 Legs	0.82	Gorilla
Tail	Backbone	0.73	Toad

Feature Space

Initial Features

- 15 binary variables
- 1 categorical variable with 6 levels (dummy coded)

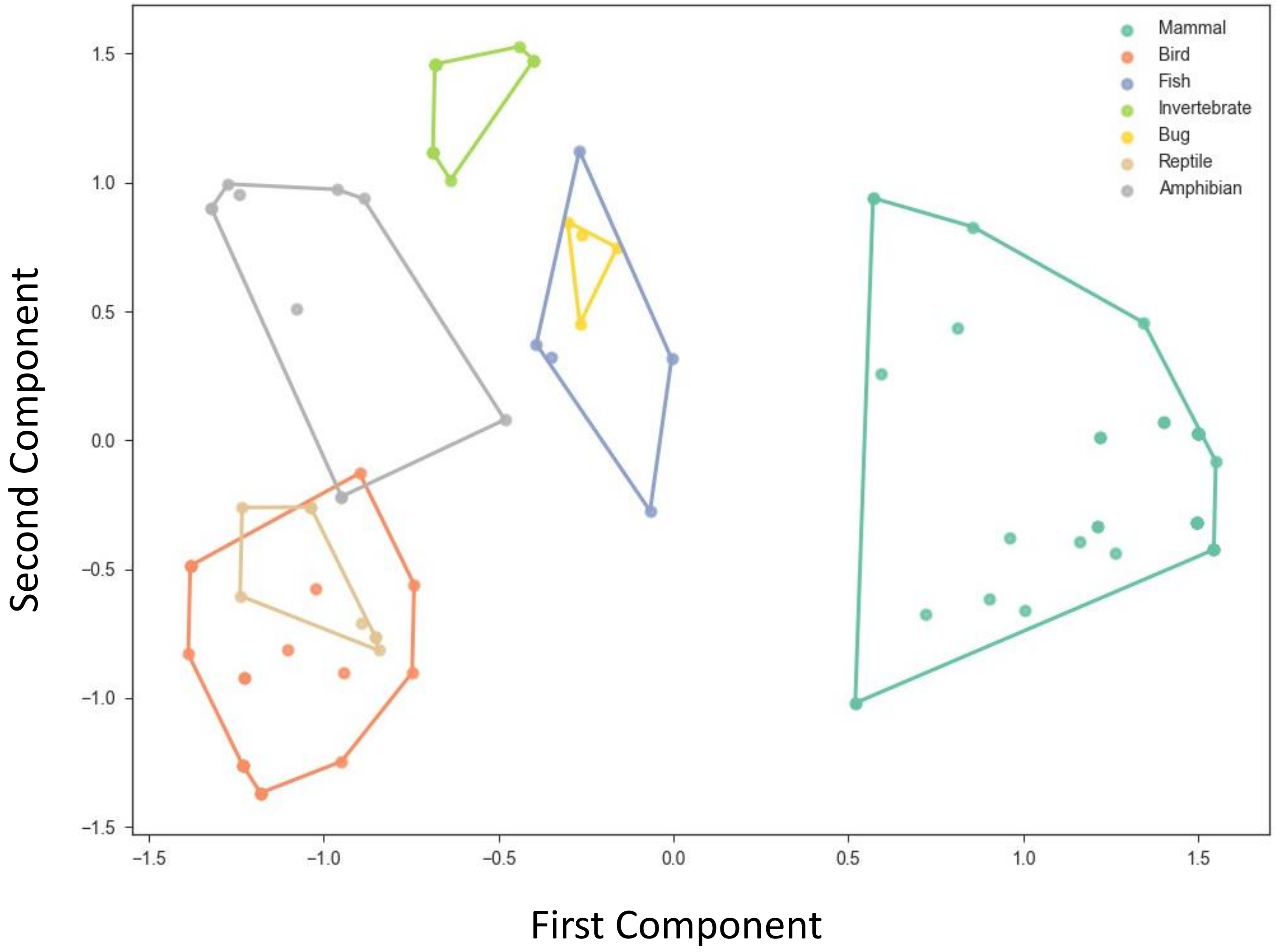
Transformed Features

- 20 binary variables

Variable 1	Variable 2	Pearson Corr.	Counterexample
Milk	Hair	0.88	Dolphin
Feathers	2 Legs	0.82	Gorilla
Tail	Backbone	0.73	Toad
Toothed	Eggs	-0.64	Haddock
6 Legs	Backbone	-0.71	Worm (neither)
Milk	Eggs	-0.94	Platypus

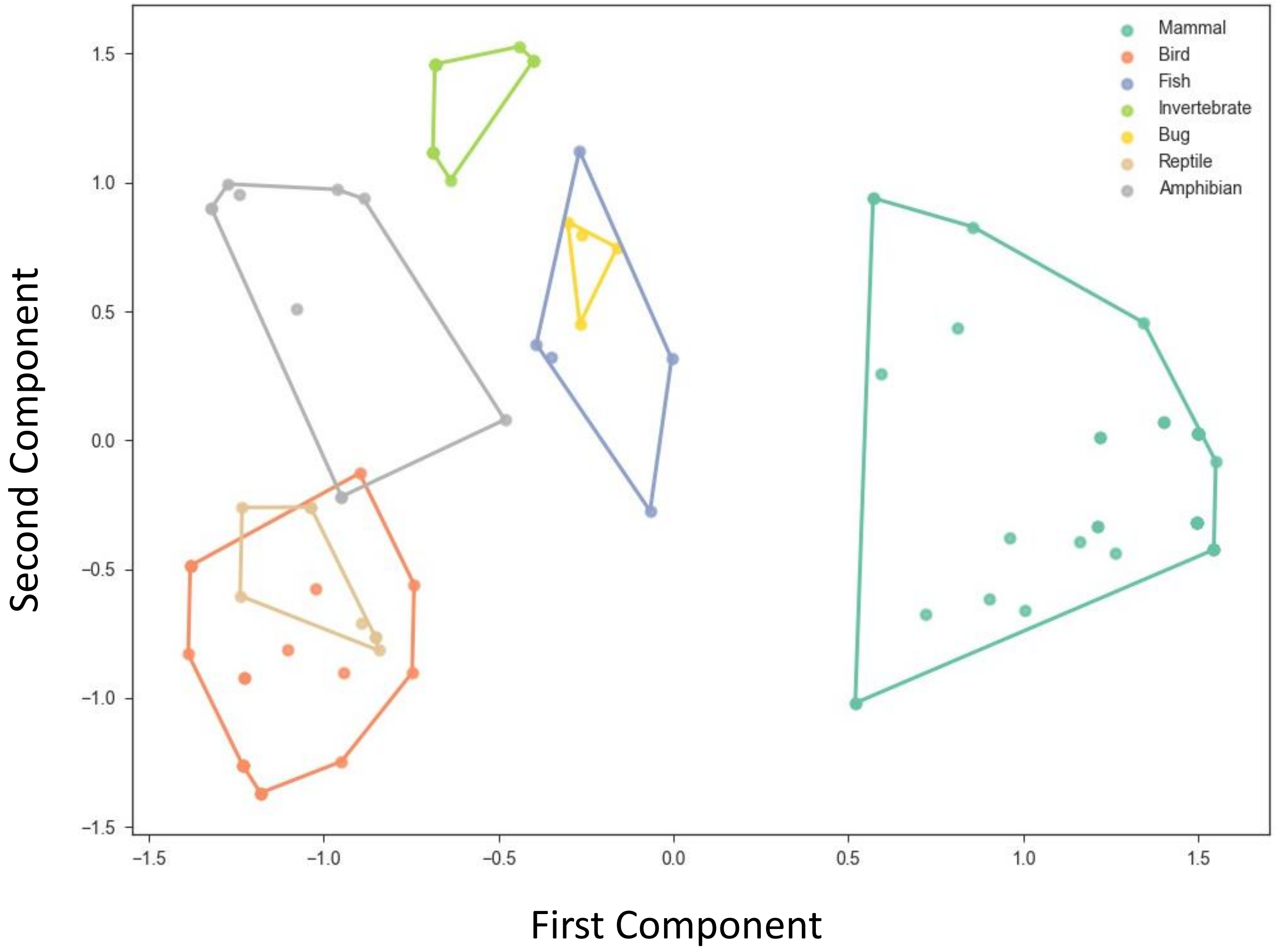
Principal Component Analysis

Ratio of Variance Explained	
PC1	35%
PC2	19%
PC3	13%
PC4	7%
PC5	5%



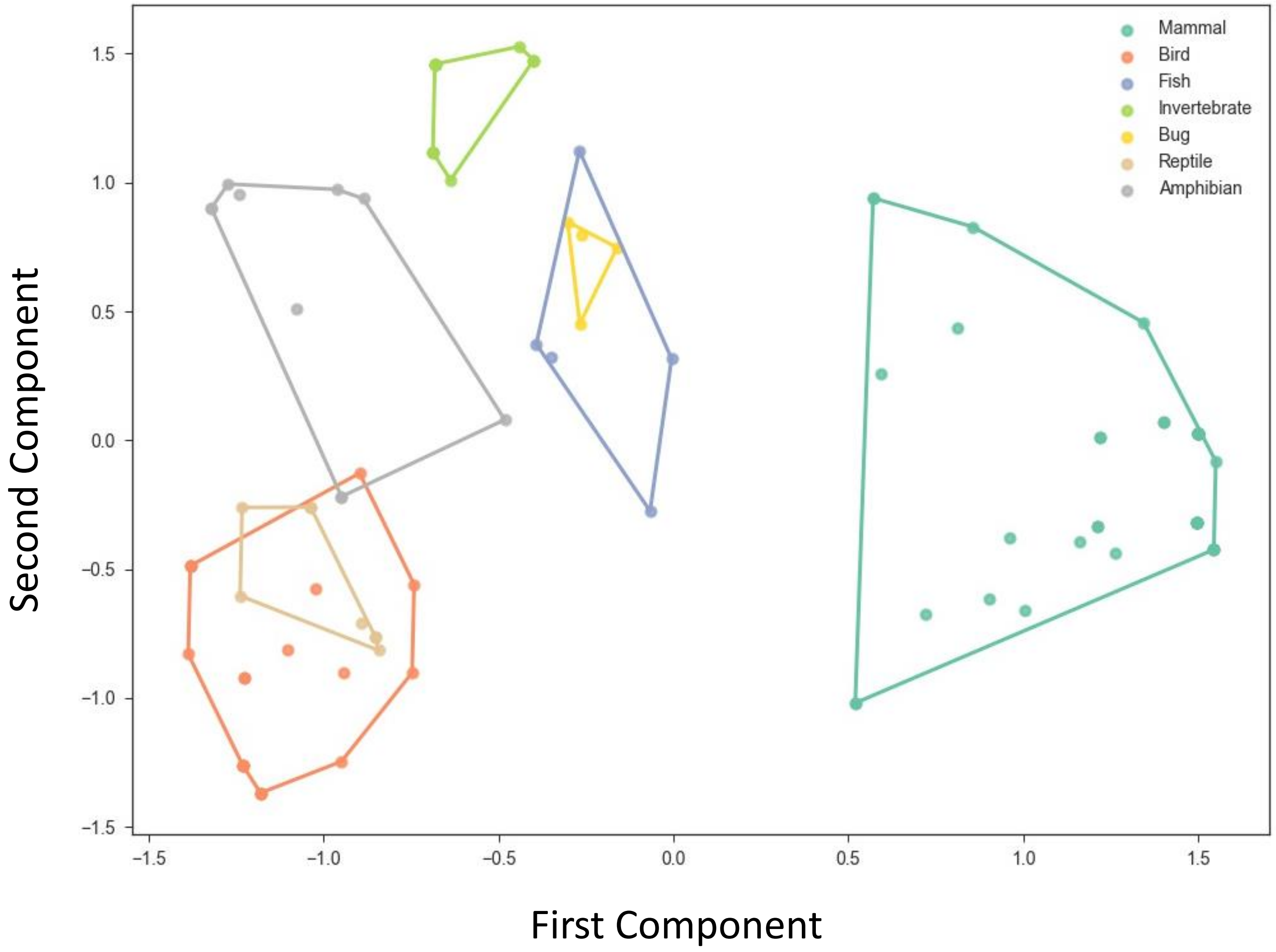
Principal Component Analysis

Variable Loadings on PC1	
Milk	0.42
4 Legs	0.35
Toothed	0.34
Eggs	-0.41



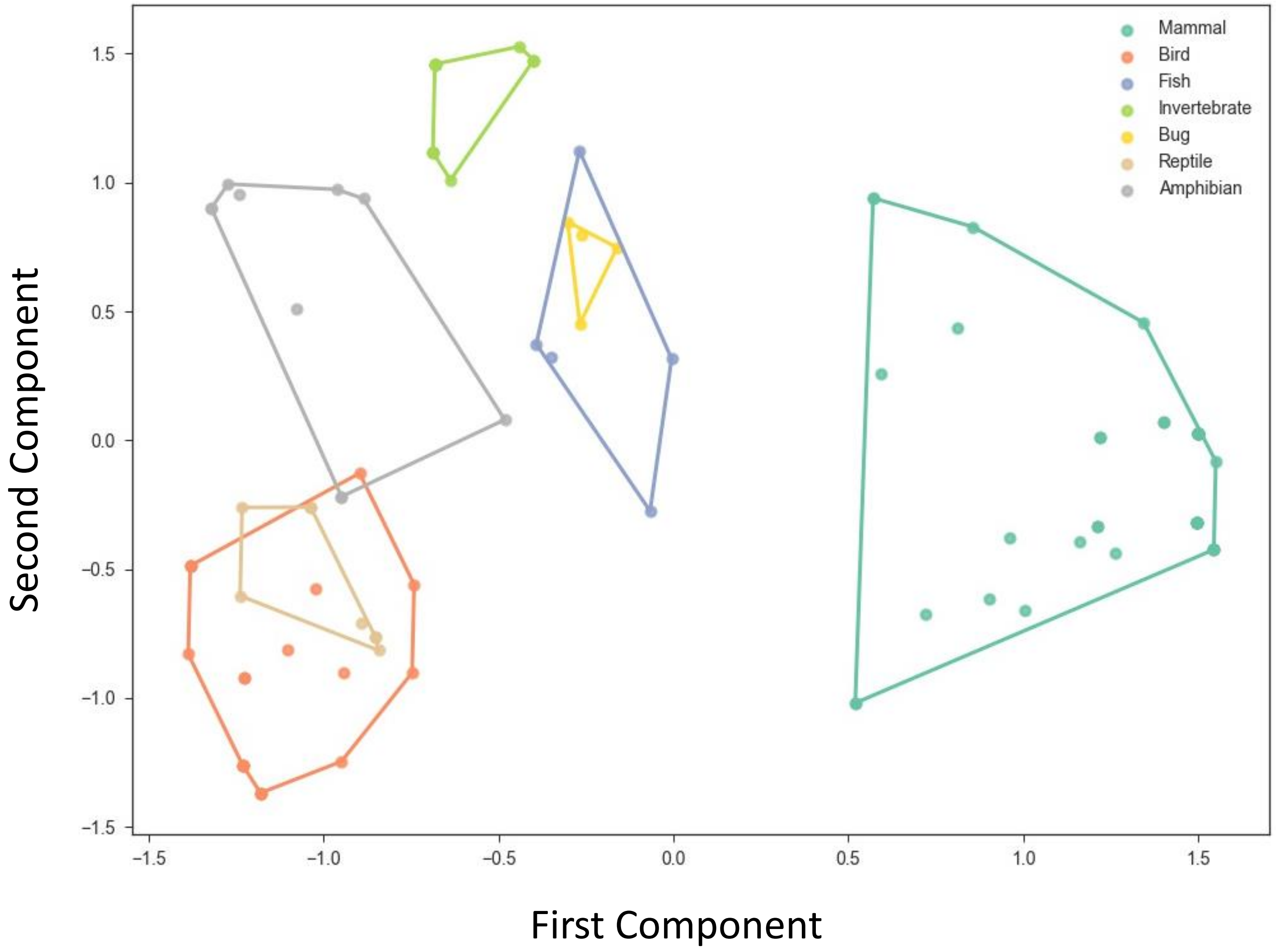
Principal Component Analysis

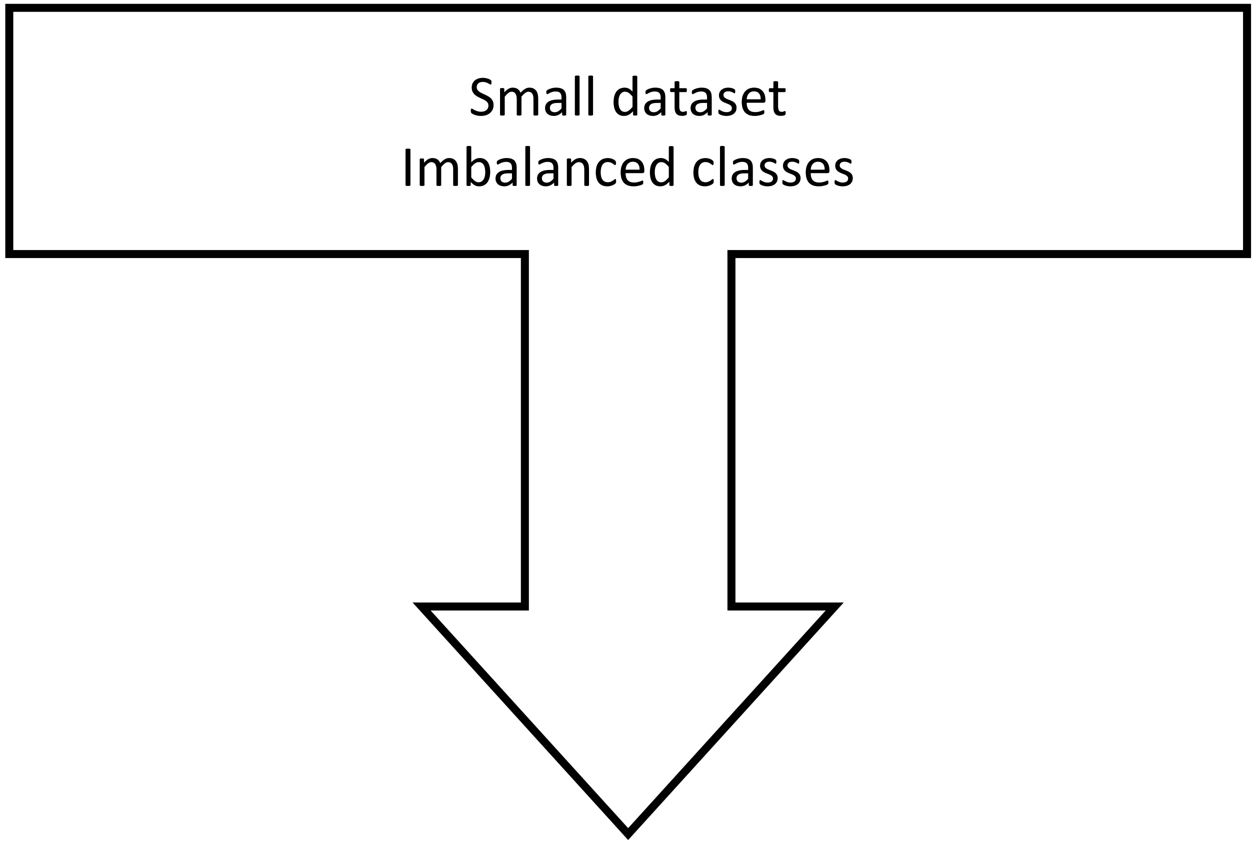
Variable Loadings on PC2	
Aquatic	0.43
Predator	0.34
Fins	0.32
Toothed	0.26
Breathes	-0.39
2 Legs	-0.35
Airborne	-0.34
Feathers	-0.29



Principal Component Analysis

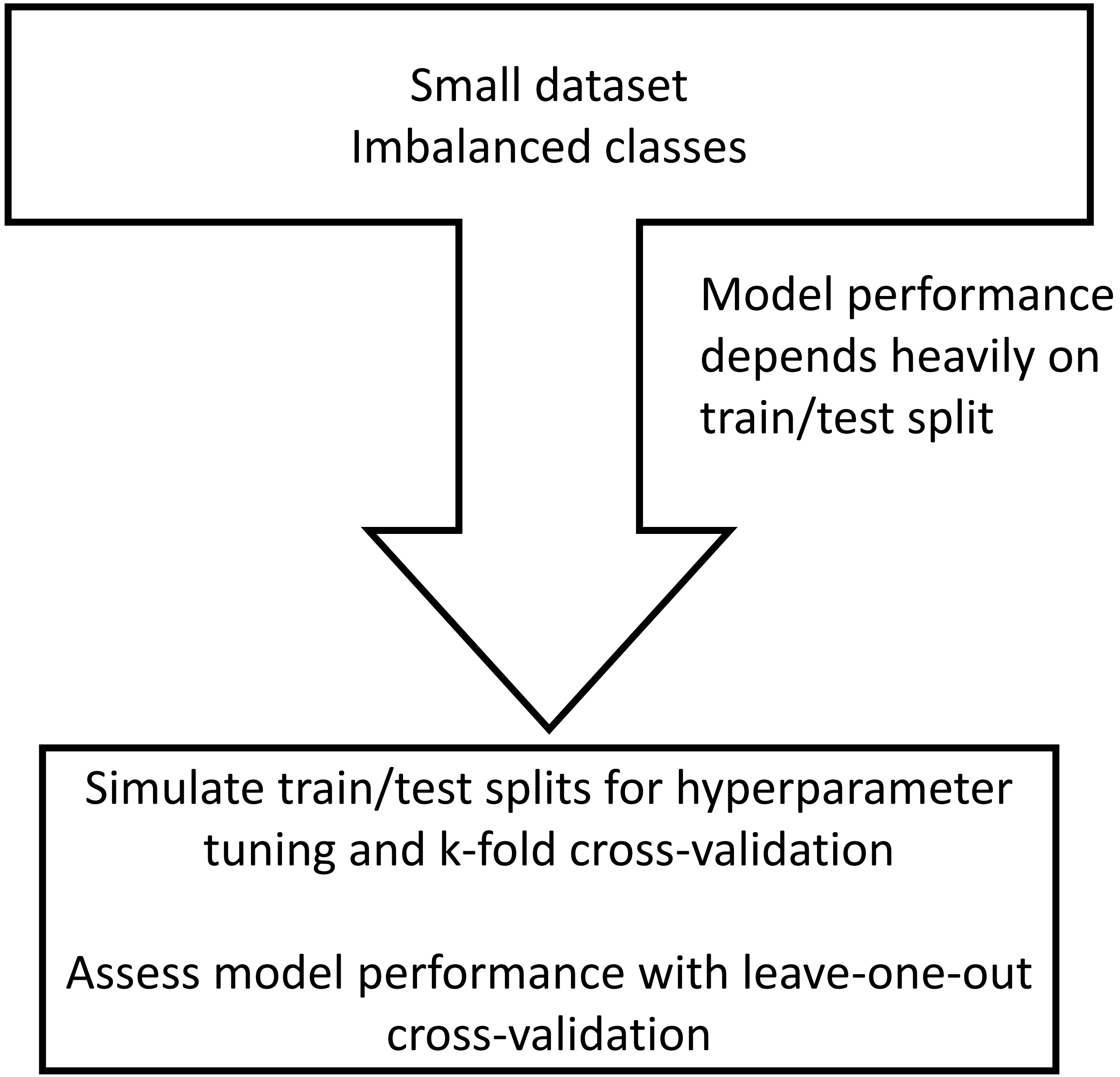
Variable Loadings on PC3	
6 Legs	0.31
Tail	-0.51
Backbone	-0.46
2 Legs	-0.36
Feathers	-0.32




Small dataset
Imbalanced classes

Small dataset
Imbalanced classes

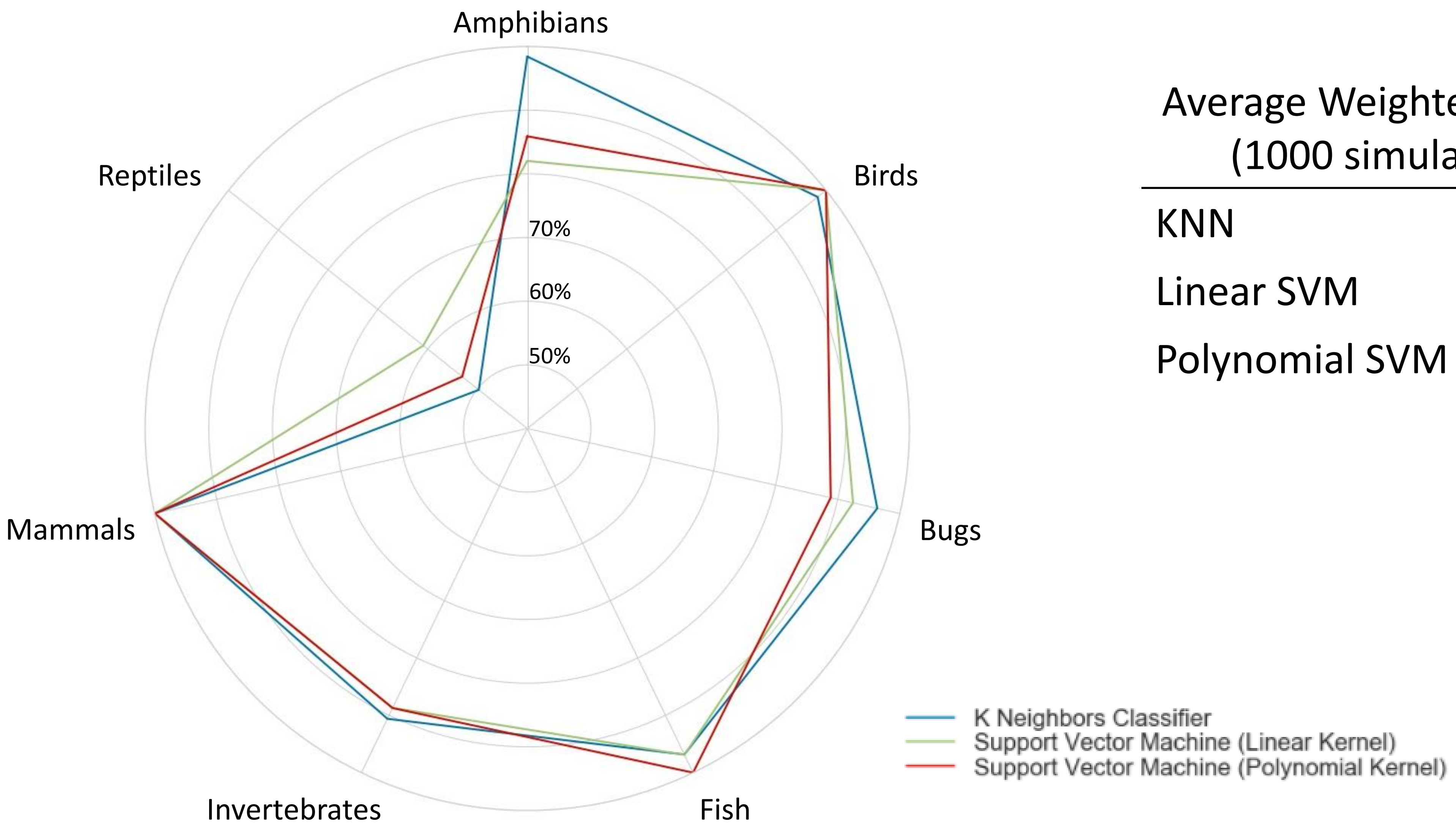
Model performance depends heavily on train/test split


Small dataset
Imbalanced classes
Model performance depends heavily on train/test split
Simulate train/test splits for hyperparameter tuning and k-fold cross-validation
Assess model performance with leave-one-out cross-validation

Models

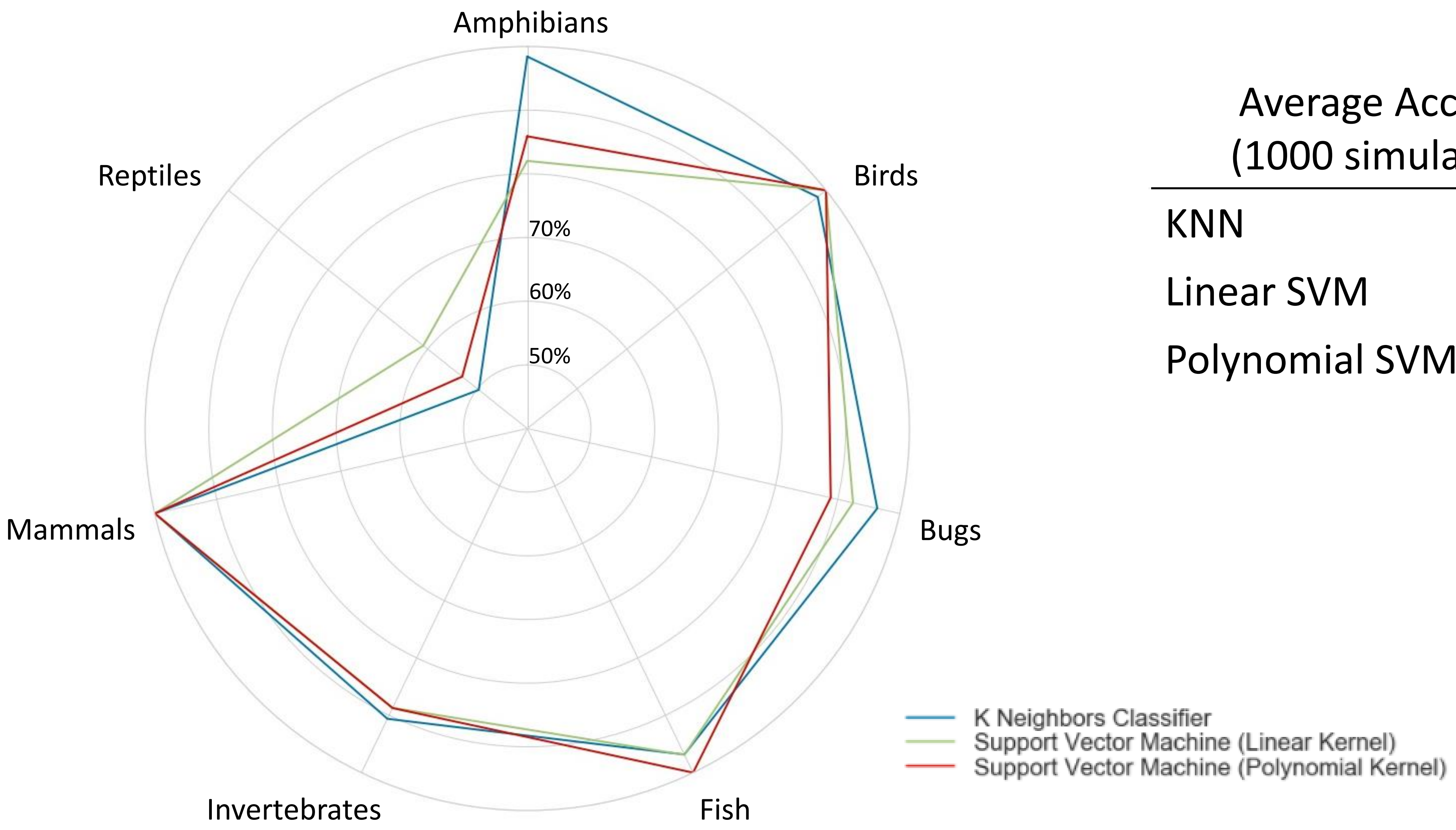
- KNN
  - Jaccard distance
  - K = 3
  - Neighbors are weighted by distance to the observation of interest.
- Support Vector Machine
  - Linear kernel, C=1
  - Polynomial kernel, C=0.01, Gamma = 1

Average F-Score by Class and Model



Average Weighted F-score (1000 simulations)	
KNN	0.960
Linear SVM	0.955
Polynomial SVM	0.955

Average F-Score by Class and Model



Average Accuracy (1000 simulations)	
KNN	96.5%
Linear SVM	96.3%
Polynomial SVM	96.5%

Leave-One-Out Cross-Validation

	Average Accuracy
KNN	97%

	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**
KNN	Scorpion	Reptile	Invertebrate
	Tortoise	Bird	Reptile

Leave-One-Out Cross-Validation

Average Accuracy	
KNN	97%
Linear SVM	97%

	Animal	Predicted Class	Actual Class
KNN	**Sea Snake**	**Fish**	**Reptile**
	Scorpion	Reptile	Invertebrate
	Tortoise	Bird	Reptile
Linear SVM	**Sea Snake**	**Fish**	**Reptile**
	Scorpion	Reptile	Invertebrate
	Newt	Reptile	Amphibian

Leave-One-Out Cross-Validation

Average Accuracy	
KNN	97%
Linear SVM	97%
Polynomial SVM	97%

	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**
KNN	Scorpion	Reptile	Invertebrate
	Tortoise	Bird	Reptile
	Sea Snake	**Fish**	**Reptile**
Linear SVM	Scorpion	Reptile	Invertebrate
	Newt	Reptile	Amphibian
	Sea Snake	**Invertebrate**	**Reptile**
Polynomial SVM	Newt	Reptile	Amphibian
	Tortoise	Invertebrate	Reptile

Recursive Feature Elimination w/ a Linear SVM

20 features

Hair
Feathers
Eggs
Milk
Airborne
Aquatic
Predator
Toothed
Backbone
Breathes
Venomous
Fins
Tail
Domestic
Catsize
2 Legs
4 Legs
5 Legs
6 Legs
8 Legs

Recursive Feature Elimination w/ a Linear SVM

20 features

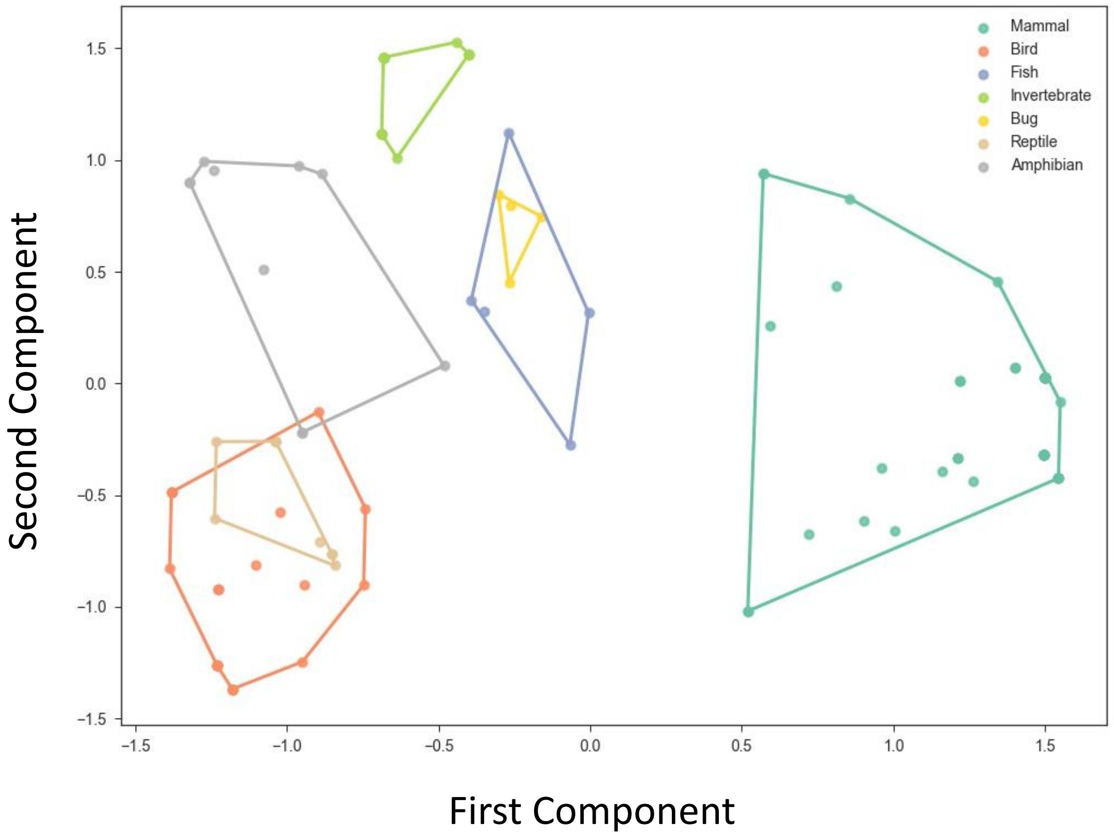
Hair
Feathers
Eggs
Milk
Airborne
Aquatic
Predator
Toothed
Backbone
Breathes
Venomous
Fins
Tail
Domestic
Catsize
2 Legs
4 Legs
5 Legs
6 Legs
8 Legs

Hair
Feathers
Eggs
Milk
Airborne
Aquatic
Toothed
Backbone
Breathes
Fins
Tail
2 Legs
4 Legs
6 Legs

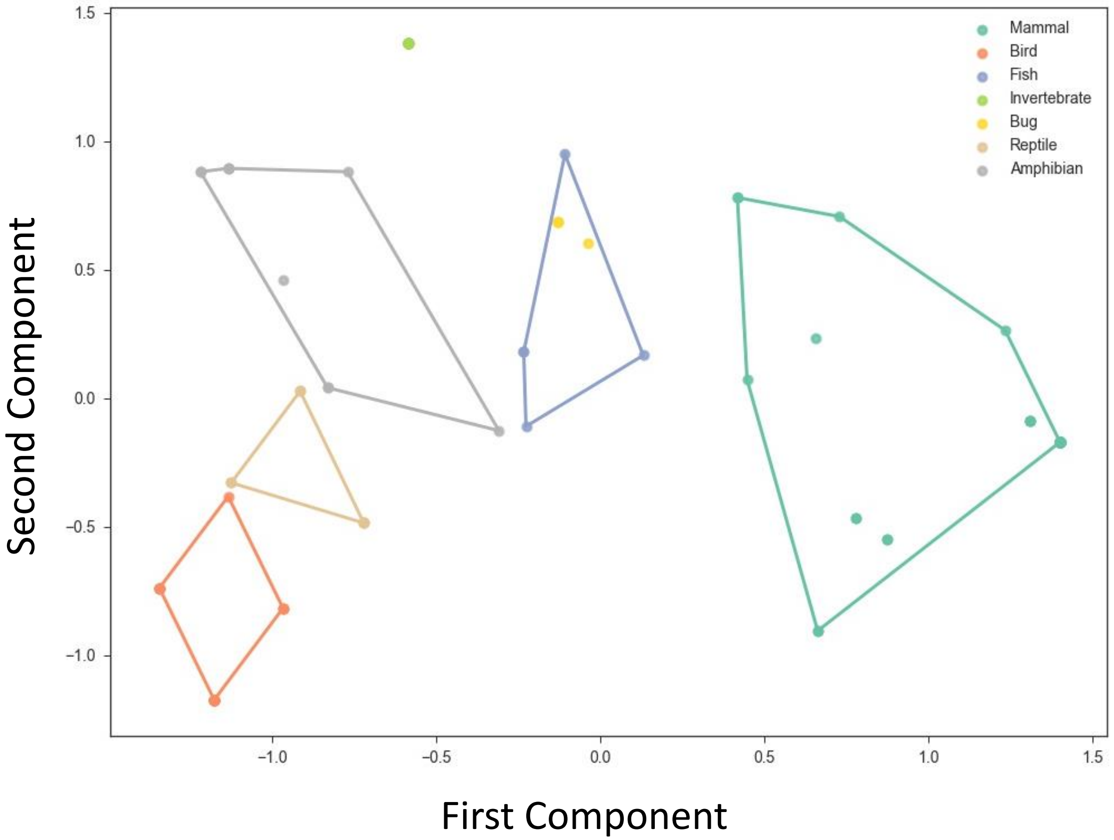
14 features

Principal Component Analysis

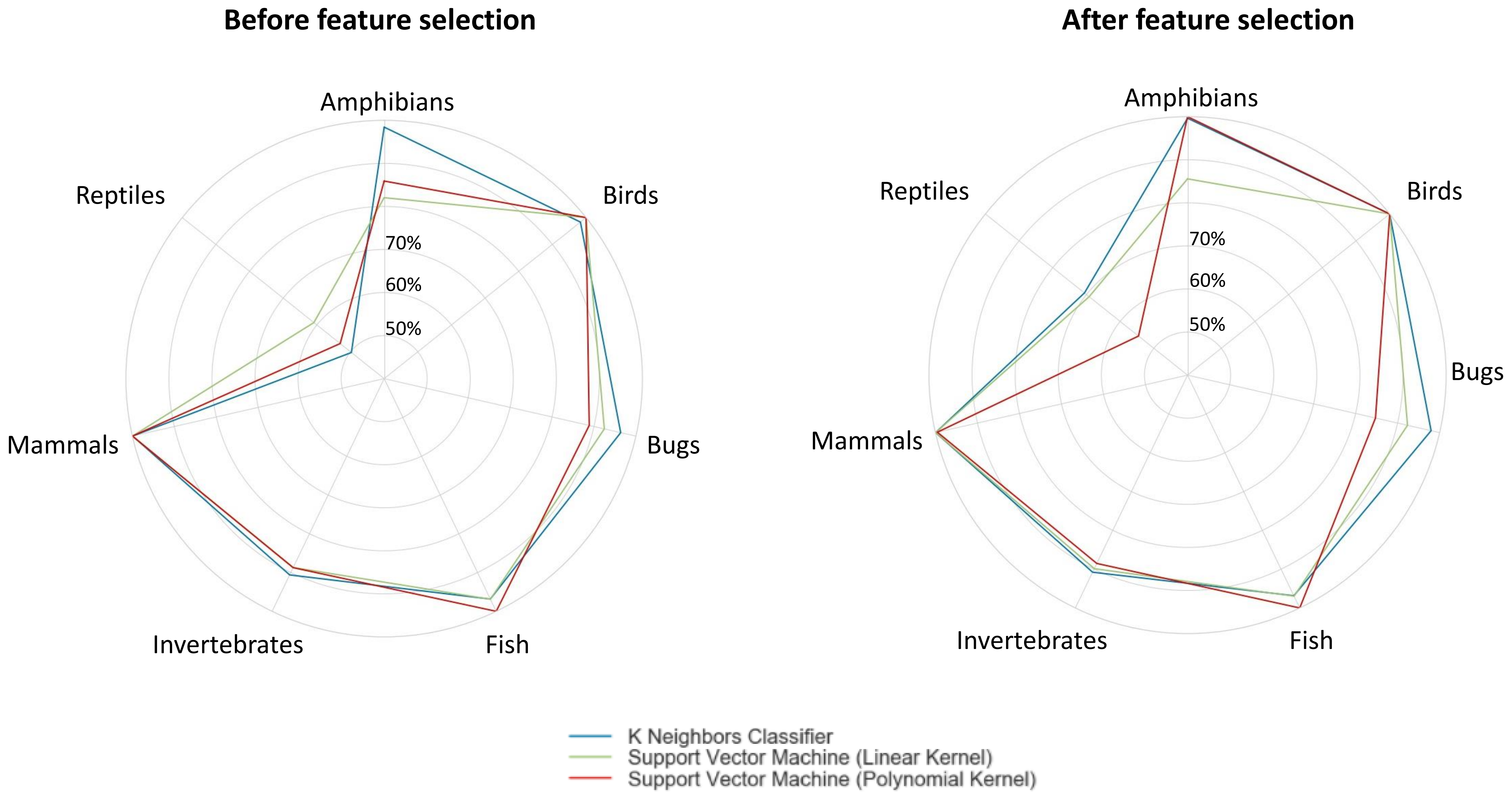
Before feature selection



After feature selection



Average F-Score by Class and Model



Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**			
KNN	Tortoise	Bird	Reptile			
	Scorpion	Reptile	Invertebrate			

Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
KNN	Tortoise	Bird	Reptile	Tortoise	Bird	Reptile
	Scorpion	Reptile	Invertebrate			

Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
KNN	**Sea Snake**	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
	Tortoise	Bird	Reptile	Tortoise	Bird	Reptile
	Scorpion	Reptile	Invertebrate			
Linear SVM	**Sea Snake**	**Fish**	**Reptile**			
	Scorpion	Reptile	Invertebrate			
	Newt	Reptile	Amphibian			

Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
KNN	Tortoise	Bird	Reptile	Tortoise	Bird	Reptile
	Scorpion	Reptile	Invertebrate			
	Sea Snake	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
Linear SVM	Scorpion	Reptile	Invertebrate	Platypus	Amphibian	Mammal
	Newt	Reptile	Amphibian			

Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
KNN	**Sea Snake**	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
	Tortoise	Bird	Reptile	Tortoise	Bird	Reptile
	Scorpion	Reptile	Invertebrate			
Linear SVM	**Sea Snake**	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
	Scorpion	Reptile	Invertebrate	Platypus	Amphibian	Mammal
	Newt	Reptile	Amphibian			
Polynomial SVM	**Sea Snake**	**Invertebrate**	**Reptile**			
	Newt	Reptile	Amphibian			
	Tortoise	Invertebrate	Reptile			

Leave-One-Out Cross-Validation

	Before Feature Elimination			After Feature Elimination		
	Animal	Predicted Class	Actual Class	Animal	Predicted Class	Actual Class
	Sea Snake	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
KNN	Tortoise	Bird	Reptile	Tortoise	Bird	Reptile
	Scorpion	Reptile	Invertebrate			
	Sea Snake	**Fish**	**Reptile**	**Sea Snake**	**Fish**	**Reptile**
Linear SVM	Scorpion	Reptile	Invertebrate	Platypus	Amphibian	Mammal
	Newt	Reptile	Amphibian			
	Sea Snake	**Invertebrate**	**Reptile**	**Sea Snake**	**Invertebrate**	**Reptile**
Polynomial SVM	Newt	Reptile	Amphibian	Platypus	Amphibian	Mammal
	Tortoise	Invertebrate	Reptile	Tortoise	Invertebrate	Reptile
				Tuatara	Amphibian	Reptile

Leave-One-Out Cross-Validation

Average Accuracy Before Feature Elimination	
KNN	97%
Linear SVM	97%
Polynomial SVM	97%

Leave-One-Out Cross-Validation

Average Accuracy Before Feature Elimination	
KNN	97%
Linear SVM	97%
Polynomial SVM	97%

Average Accuracy After Feature Elimination	
KNN	98%
Linear SVM	98%
Polynomial SVM	96%

Why can’t we get the sea snake right?!

- The reptile data is too small and too diverse





So what makes a sea snake anyways?

- Live the majority of their lives in the water
- Have developed tails for better swimming
- Some species lay eggs on land, but the majority have live births
- Need air regularly as they do not have gills

So what makes a sea snake anyways?

- Live the majority of their lives in the water
- Have developed tails for better swimming
- Some species lay eggs on land, but the majority have live births
- **Need air regularly as they do not have gills**

Wait a second…

hair	feathers	eggs	milk	airborne	aquatic	toothed	backbone	breathes	fins	tail	2 legs	4 legs	6 legs
0	0	0	0	0	1	1	1	0	0	1	0	0	0

Wait a second…

hair	feathers	eggs	milk	airborne	aquatic	toothed	backbone	breathes	fins	tail	2 legs	4 legs	6 legs
0	0	0	0	0	1	1	1	$\boxed{0}$	0	1	0	0	0

The dataset is WRONG!


Big data gets a lot of attention these days,
but small data presents a unique challenge.


Big data gets a lot of attention these days,
but small data presents a unique challenge.
WATCH OUT FOR SEA SNAKES!

Appendix

Problem Description

Given 101 observations of 20 binary characteristics, can we predict an animals classification in the Animalia kingdom? If so, how well can we do it, and how can we ensure model stability with such a small amount of data?

Data Cleaning

- Renaming frog labels as ‘frog1’ and ‘frog2’

Feature Engineering

- Encoding the *legs* variable
- Recursive feature elimination with a support vector machine

Python Tools and Packages

Data Manipulation

pandas, numpy

Graphing

matplotlib, seaborn, yellowbrick, pylab

Modeling

scikit-learn, scipy