Species Distribution Modeling

Julie Lapidus
Scripps College ‘11

Eli Moss
Brown University ‘11

Objective

- To characterize the performance of both multiple response and single response machine learning algorithms for multiple datasets

Single Response vs. Multiple Response Algorithms

- 1. Single-response models
  - input: covariates
  - output: prediction for a single species
- 2. Multiple-response models
  - input: covariates
  - output: simultaneous predictions for all species

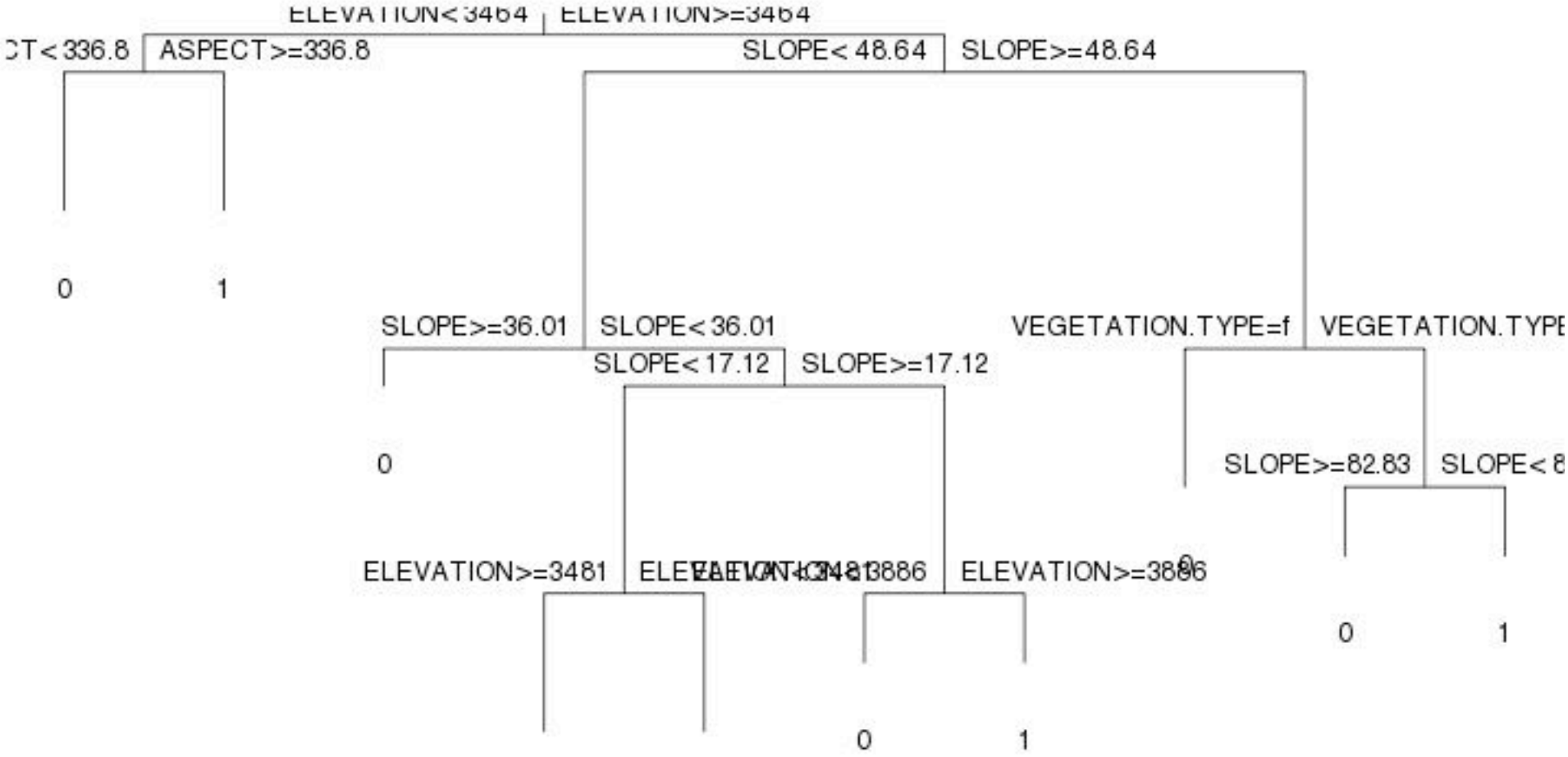
Single vs. Multiple Response

- Single: Data learning only needs to predict one response at a time.
  - Predictions informed solely by patterns in covariates
- Multiple: Fit of data attempts to account for all responses simultaneously.
  - Prediction of rare moths might be helpfully influenced by patterns in others (only if they covary)

Single Response

- Elastic net logistic regression: fit the data by weighting each covariate within a linear equation.
  - Control overfitting by penalizing weights
- Decision trees: make successive splits on covariates to arrive at an output
  - Control overfitting by reducing size of tree

Single Response



Multiple Response

- Multivariate Decision trees: splits must predict all species, are chosen according to best overall correlation
  - Control overfitting by reducing size of tree
- Single-Hidden-Layer Neural Network: Non-linear statistical data modeling tool inspired by human neural networks
  - Control overfitting by selecting best number of training iterations

Moth Data

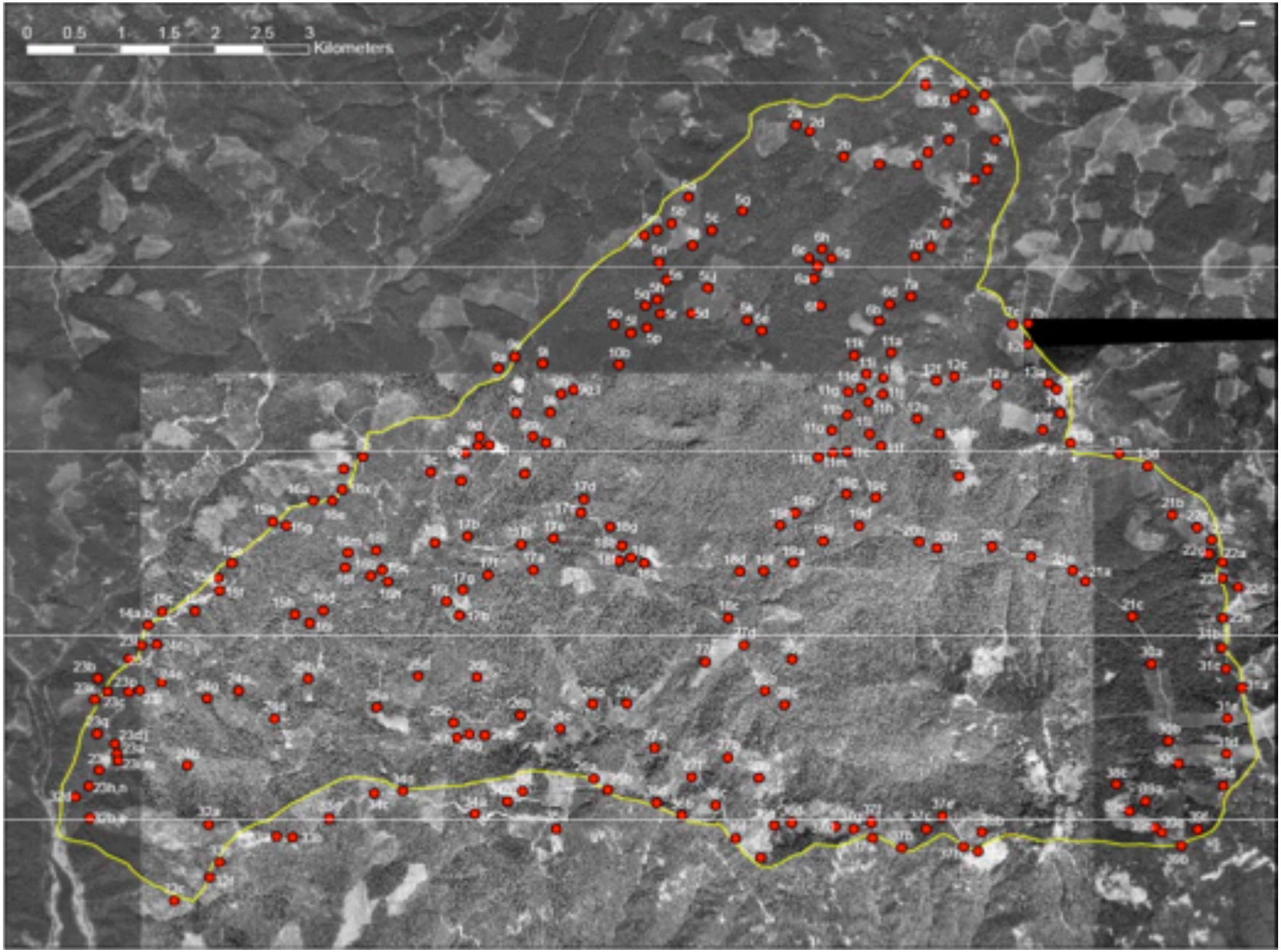
- Jeff Miller, collected ‘86-’08
- H.J. Andrews Experimental Forest
- 606 species
- 4 covariates
- 256 traps

Environmental Covariates

- Slope: percent grade 1 - 105
- Aspect: degrees 2 - 358
- Elevation: meters 1437 - 5007
- Vegetation Type: closed forest, meadow, cut 72-77, open forest, shrub/very open forest
  - represented as numbers when modeled

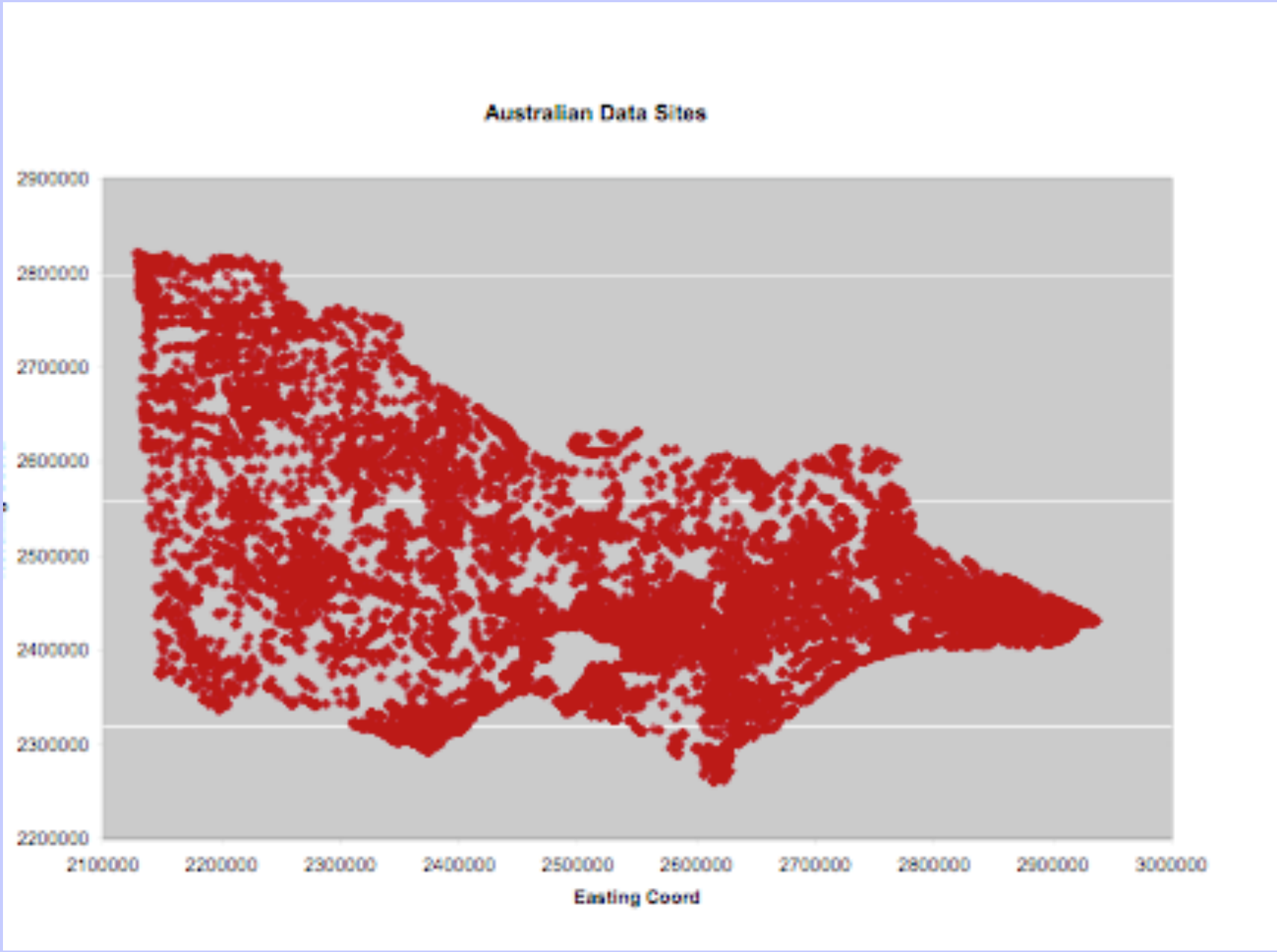
Issues with Moth Dataset

- Not uniformly sampled (roads)
- > 1/2 of the included species occurred fewer than 18 times over the course of the entire 23-year trapping period
- 1/6 of the species account for over half of the recorded moth occurrences
- Euclidian transformation of slope and aspect necessary
- Different moth species more prevalent each year


0 0.5 1 1.5 2 2.5 3
Kilometers

Australian Plants Dataset

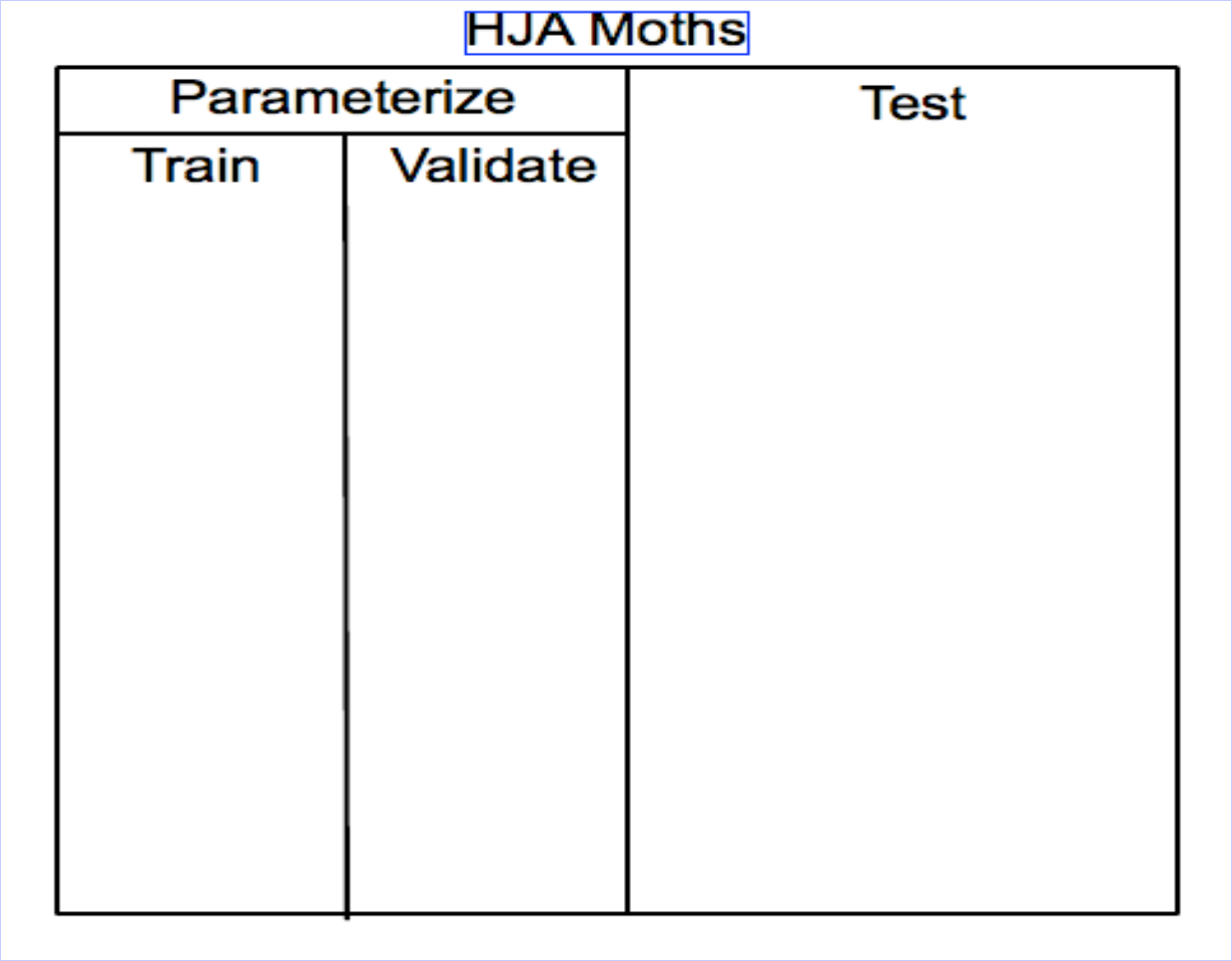
- Victoria, Australia
- 5000 plant species
- Arthur Rylah Institute (Melbourne, Australia)
- Subset of 100 most abundant plant species
- 15,328 sites
- 81 covariates


Australian Data Sites
2900000
2800000
2700000
2600000
2500000
2400000
2300000
2200000
2100000
2200000
2300000
2400000
2500000
2600000
2700000
2800000
2900000
3000000
Easting Coord

Data Format

SiteID	Covariate N (e.g. temperature)	Species 1	Species n
A	x	0	1
B	y	1	1
C	z	1	0

Subsetting the Data



Cohen's Kappa

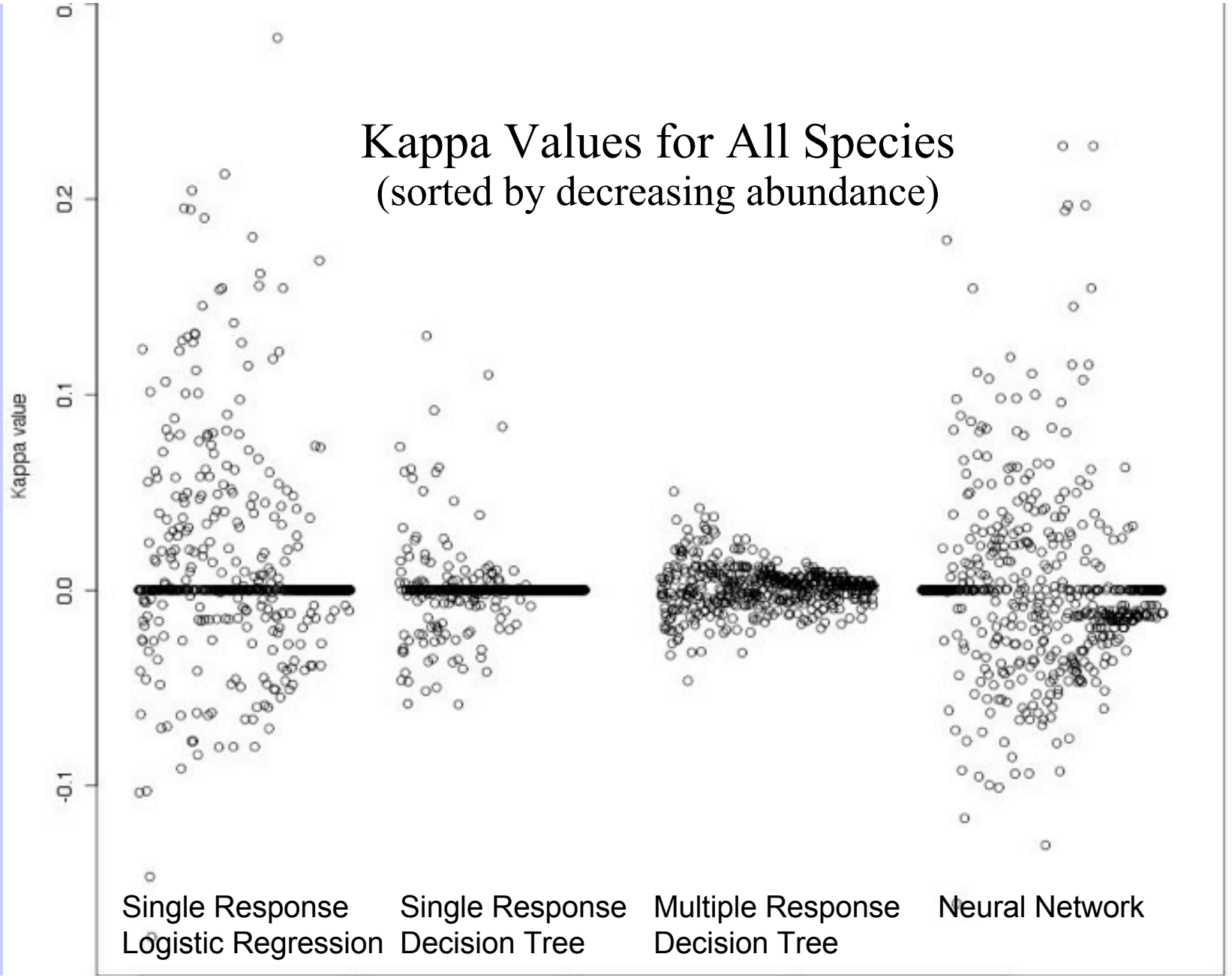
- Statistical performance measurement used

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

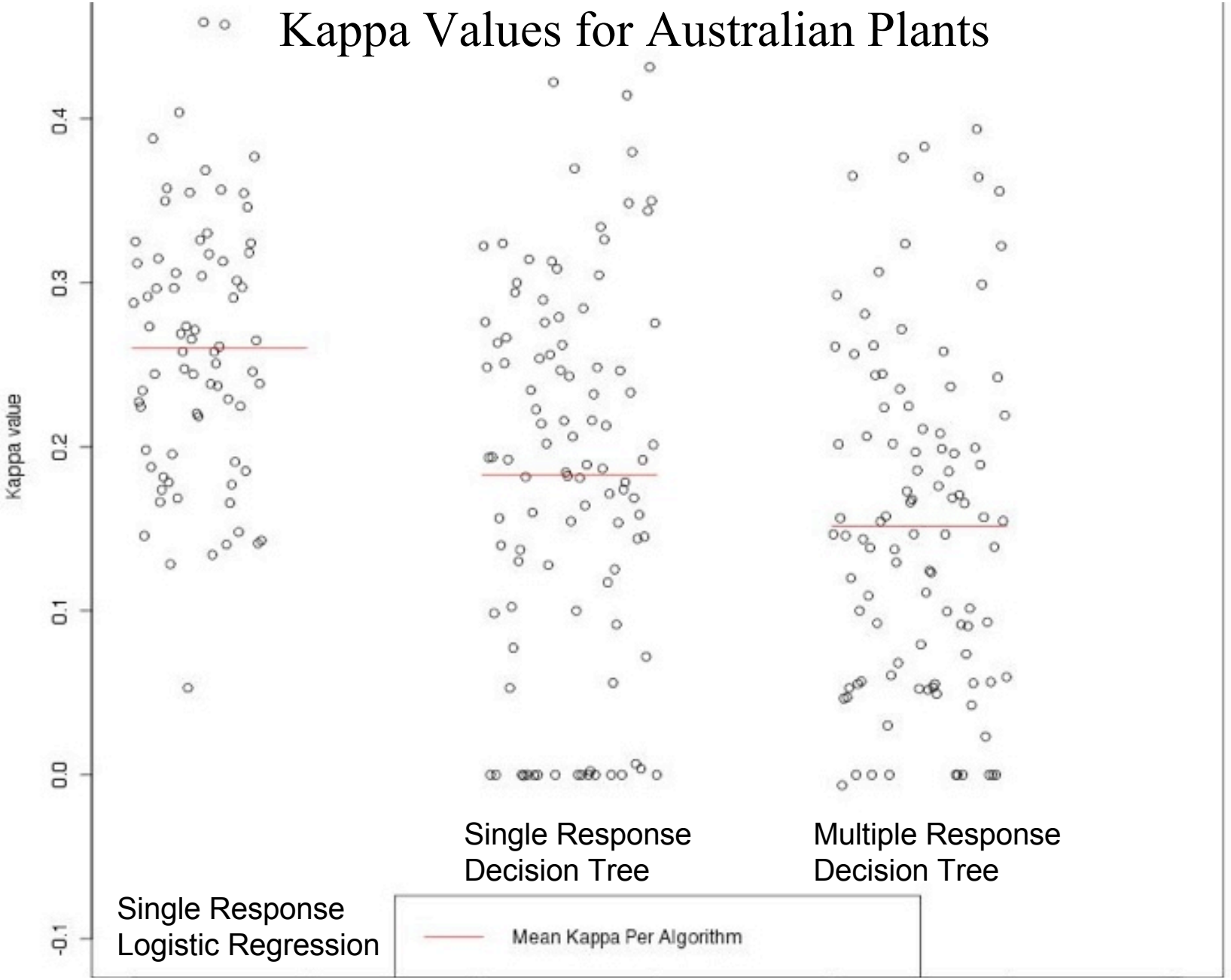
- Pr(a) is relative observed agreement among raters
- Pr(e) is hypothetical probability oof chance agreement

Value	Interpretation
1	Complete agreement
0	No agreement beyond chance
<0	Worse than random guessing

Kappa Values for All Species
(sorted by decreasing abundance)

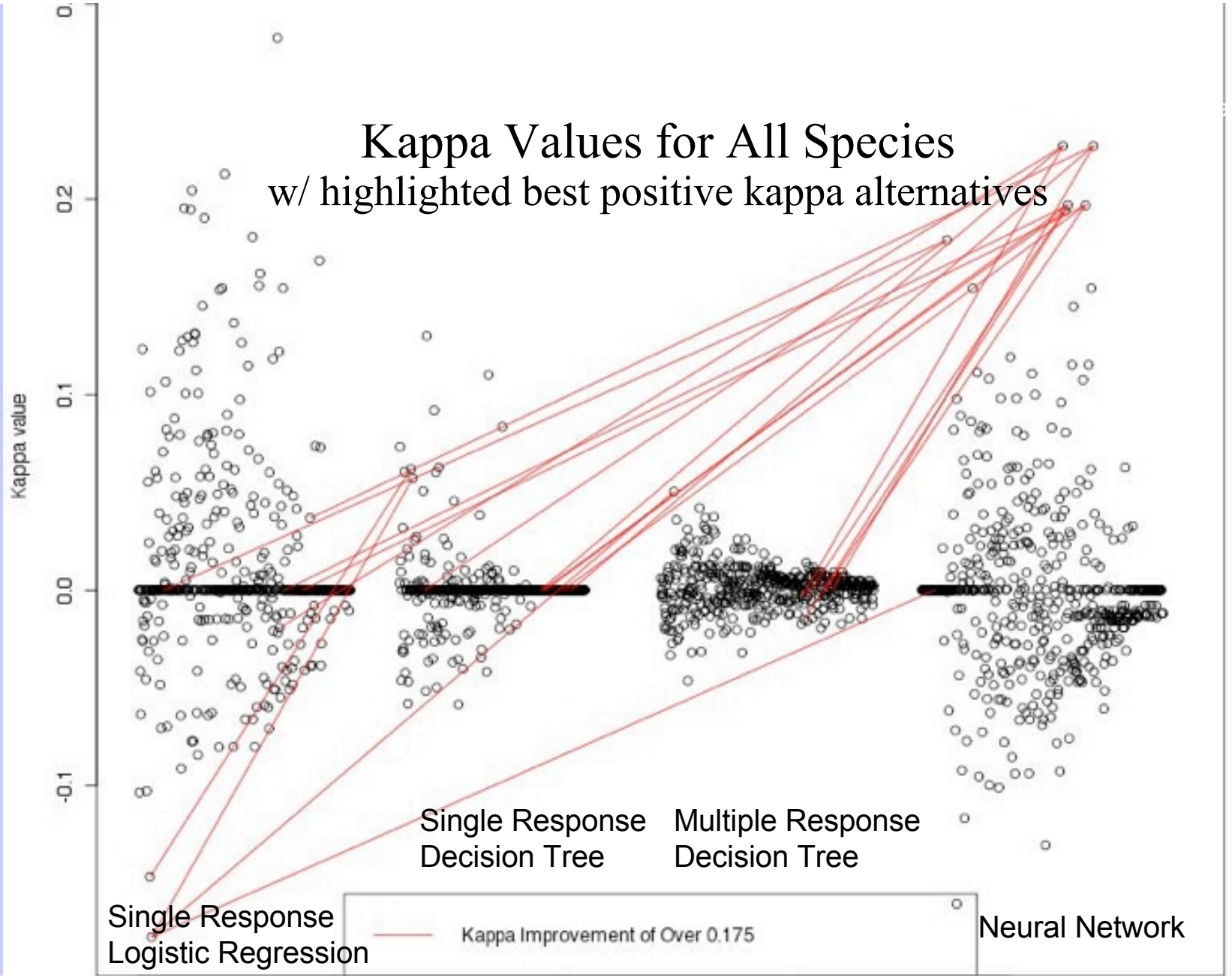


Kappa Values for Australian Plants



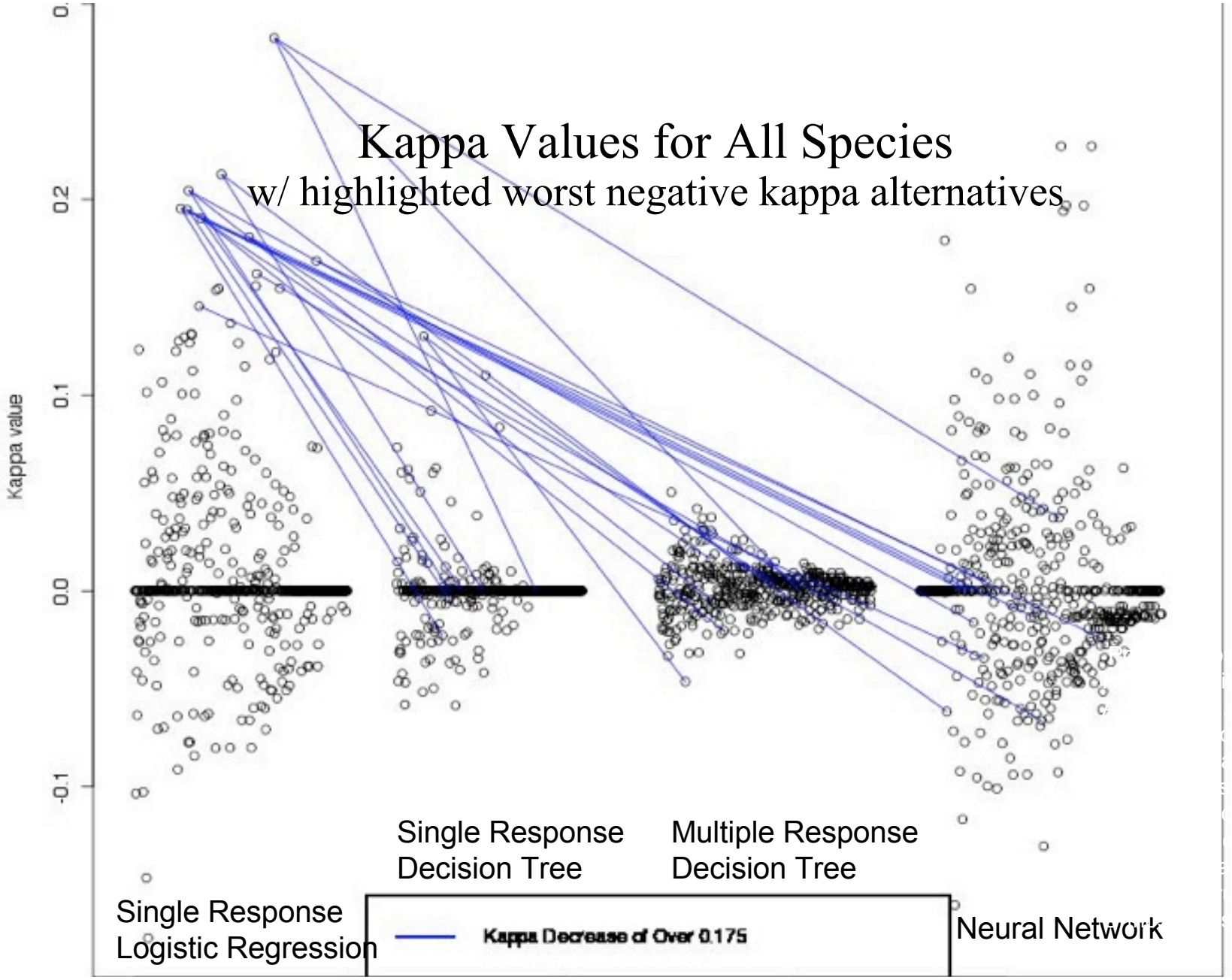
Improving Predictions

- Difficult prediction problem
  - few samples for many of the moths
  - Few covariates
  - Coarse vegetation type covariate
  - non-continuous aspect covariate
- Creative use of our tools can yield better predictions...
  - algorithms perform best on different ranges of abundance/covariate values

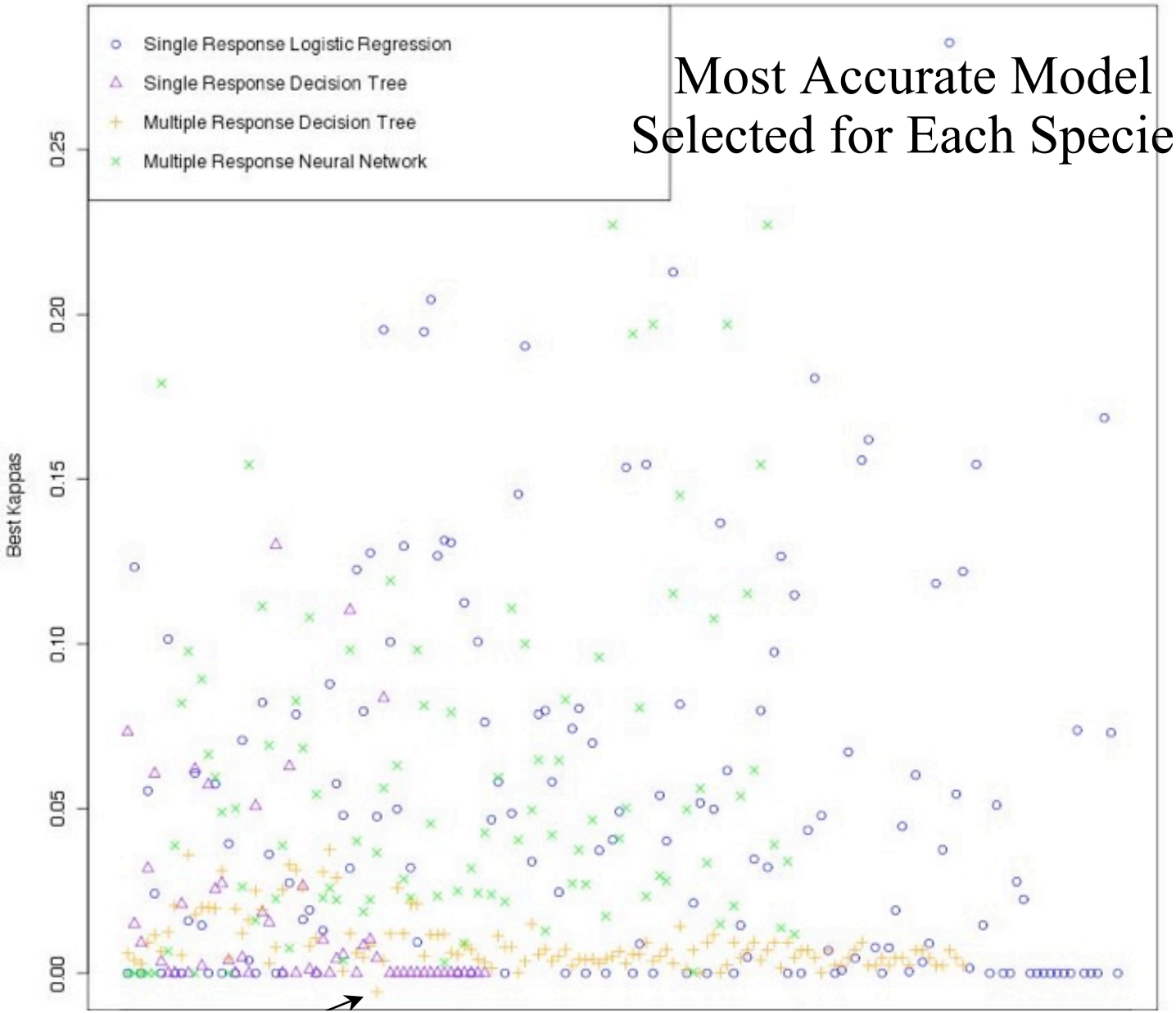

Kappa Values for All Species
w/ highlighted best positive kappa alternatives
Kappa value
0.2
0.1
0.0
-0.1
Single Response
Decision Tree
Multiple Response
Decision Tree
Single Response
Logistic Regression
Kappa Improvement of Over 0.175
Neural Network

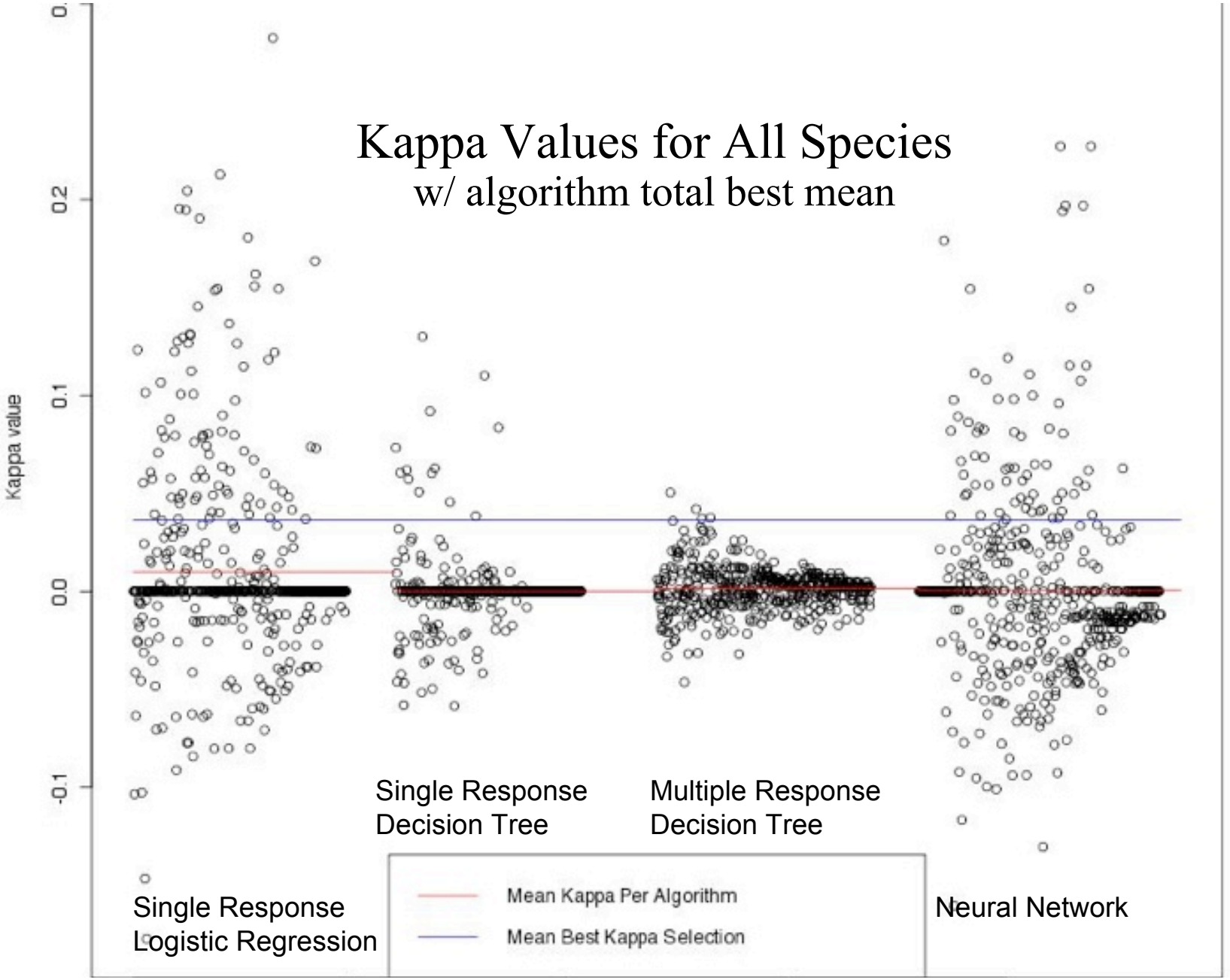
Kappa Values for All Species

w/ highlighted worst negative kappa alternatives



Most Accurate Model Selected for Each Species




Kappa Values for All Species
w/ algorithm total best mean
Kappa value
0.2
0.1
0.0
-0.1
Single Response
Logistic Regression
Single Response
Decision Tree
Multiple Response
Decision Tree
Neural Network
Mean Kappa Per Algorithm
Mean Best Kappa Selection

Summary of Results

- Tested four algorithms for two datasets
- We have established a comparative baseline for predictive performance with the Australian plant dataset
- The moth dataset poses a considerable problem to modeling
  - noisy occurrence patterns
  - thinly sampled occurrence for 5/6 species
- Integration of models may improve prediction accuracy
  - algorithm selection based on abundance, covariate proportions for each species

Further Research

- Data Preprocessing
  - group moths by habitat preferences
  - perform euclidean transformation on slope/aspect
  - continuous vegetation type
- Prediction of moth species occurrence across HJA with environment grid
- More advanced algorithms
  - our basic methods will serve as a baseline for comparison

Acknowledgements

- We would like to thank:
  - Dr. Dietterich, CS Professor
  - Dr. Wong, CS Professor
  - Steven Highland, Geosciences PhD candidate
  - Julia Jones, Geosciences Professor
  - Arwen Lettkeman, CS PhD candidate
  - Paul Wilkins, CS PhD candidate
  - Rebecca Hutchinson, CS Post-doc