Random forest and predictive phylogeography

SSB 2018

June 1 Friday 1pm-5pm

Anahí Espíndola, Megan Smith, Megan Ruffley, Tara Pelletier

Today

- Predictive modeling
- Random forest
- Examples in phylogeography
- Hands-on tutorial
- Assessing model accuracy
- Hands-on tutorial
- Predicting and variable importance
- Hands-on tutorial

This is the variable we are trying to predict **(response variable)**.

Predictive Modeling

Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder Woman	9	Patty Jenkins	Warner Bros	action, adventure, fantasy	god, wonder woman, mission	7.6	76
Logan	7	James Mangold	20th Century Fox	action, drama, sci-fi	x-men, marvel, superhero	8.1	77
Zootopia	9	Byron Howard	Walt Disney	animation, adventure, comedy	fox, police, con artist	8	78
Captain America: Civil War	8	Anthony Russo	Marvel Studios	action, adventure, sci-fi	marvel, comic book, superhero	7.8	75
Beauty and the Beast	6	Bill Condon	Mandeville	family, fantasy, musical	beast, fairy tale, disney	7.3	65
Moana	7	Ron Clements	Walt Disney	animation, adventure, comedy	island, ocean, polynesia	7.6	81
Rogue One: A Star Wars Story	9	Gareth Edwards	Lucasfilm	action, adventure, sci-fi	loss of loved one, star wars, female criminal	7.8	65
Deadpool	7	Tim Miller	20th Century Fox	action, adventure, comedy	breaking forth wall, self healing, comic book	8	65
Finding Dory	9	Andrew Stanton	Pixar	animation, adventure, comedy	fish, ocean, whale	7.4	77
Ghostbusters	10	Paul Feig	Columbia	action, comedy, fantasy	vomit, remake, feminism	5.3	60
Suicude Squad	6	David Ayer	Atlas Entertainment	action, adventure, fantasy	comic book, suicide squad, father daugher relationship	6.1	40
Ant-Man	8	Payten Reed	Gary Sanchez	action, adventure, comedy	heist, sabotage, vault	7.3	64
Jurassic World	9	Colin Trevorrow	Universal	action, adventure, sci-fi	dinosaur, jurassic park, velociraptor	7	59

Predictive Modeling

These are our **predictor variables**.

Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder Woman	9	Patty Jenkins	Warner Bros	action, adventure, fantasy	god, wonder woman, mission	7.6	76
Logan	7	James Mangold	20th Century Fox	action, drama, sci-fi	x-men, marvel, superhero	8.1	77
Zootopia	9	Byron Howard	Walt Disney	animation, adventure, comedy	fox, police, con artist	8	78
Captain America: Civil War	8	Anthony Russo	Marvel Studios	action, adventure, sci-fi	marvel, comic book, superhero	7.8	75
Beauty and the Beast	6	Bill Condon	Mandeville	family, fantasy, musical	beast, fairy tale, disney	7.3	65
Moana	7	Ron Clements	Walt Disney	animation, adventure, comedy	island, ocean, polynesia	7.6	81
Rogue One: A Star Wars Story	9	Gareth Edwards	Lucasfilm	action, adventure, sci-fi	loss of loved one, star wars, female criminal	7.8	65
Deadpool	7	Tim Miller	20th Century Fox	action, adventure, comedy	breaking forth wall, self healing, comic book	8	65
Finding Dory	9	Andrew Stanton	Pixar	animation, adventure, comedy	fish, ocean, whale	7.4	77
Ghostbusters	10	Paul Feig	Columbia	action, comedy, fantasy	vomit, remake, feminism	5.3	60
Suicude Squad	6	David Ayer	Atlas Entertainment	action, adventure, fantasy	comic book, suicide squad, father daugher relationship	6.1	40
Ant-Man	8	Payten Reed	Gary Sanchez	action, adventure, comedy	heist, sabotage, vault	7.3	64
Jurassic World	9	Colin Trevorrow	Universal	action, adventure, sci-fi	dinosaur, jurassic park, velociraptor	7	59

Predictive Modeling

Movie	My score	Director	Production Co	1st three genres	1st three plot keywords	IMDB score	Meta-score
Star Wars: The Last Jedi	10	Rian Johnson	Lucasfilm	action, adventure, fantasy	wisecrack humor, deception, betrayal	7.5	85
Wonder Woman	9	Patty Jenkins	Warner Bros	action, adventure, fantasy	god, wonder woman, mission	7.6	76
Logan	7	James Mangold	20th Century Fox	action, drama, sci-fi	x-men, marvel, superhero	8.1	77
Zootopia	9	Byron	Walt Disney	animation, adventure,	fox, police, con artist	8	78

How would I score Black Panther based on this set of predictor variables using a set of know movie scores?

[illegible]	[illegible]	[illegible]	Fox	comedy	healing, comic book	[illegible]	[illegible]
Finding Dory	9	Andrew Stanton	Pixar	animation, adventure, comedy	fish, ocean, whale	7.4	77
Ghostbusters	10	Paul Feig	Columbia	action, comedy, fantasy	vomit, remake, feminism	5.3	60
Suicude Squad	6	David Ayer	Atlas Entertainment	action, adventure, fantasy	comic book, suicide squad, father daugher relationship	6.1	40
Ant-Man	8	Payten Reed	Gary Sanchez	action, adventure, comedy	heist, sabotage, vault	7.3	64
Jurassic World	9	Colin Trevorrow	Universal	action, adventure, sci-fi	dinosaur, jurassic park, velociraptor	7	59

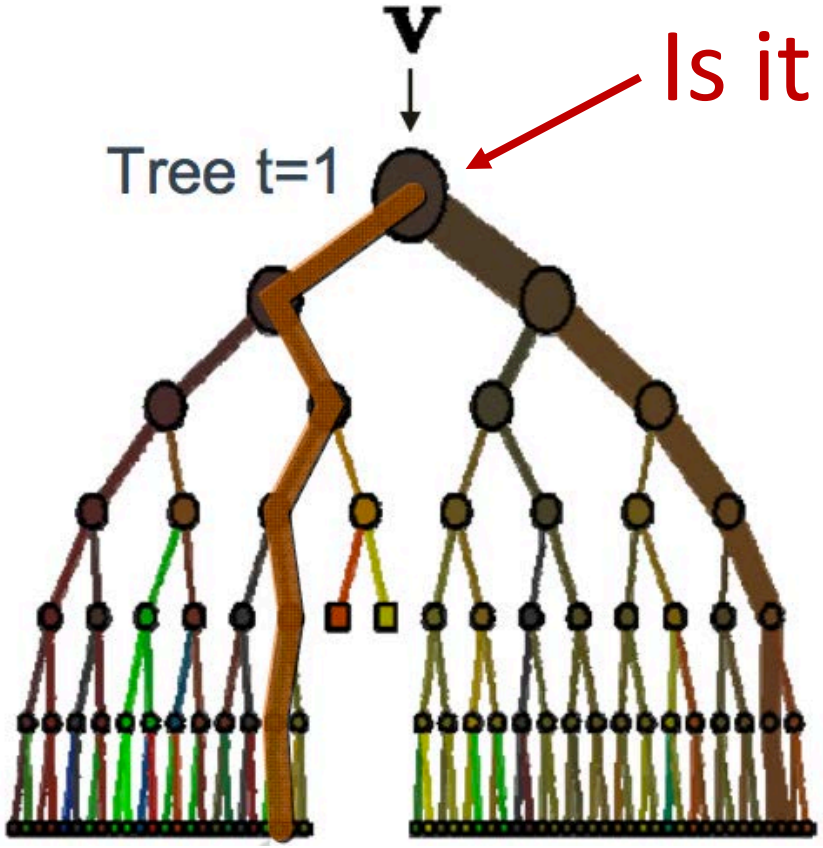
Predictive Phylogeography

- Using large datasets that include various types of data (genetic, geographic, environmental, morphological) to build a model (**training data**), to then make phylogeographic predictions on species, populations, or individuals with an unknown response (**test data**)
- Examples in a few minutes by:
  - Megan Smith
  - Megan Ruffley

Random Forest

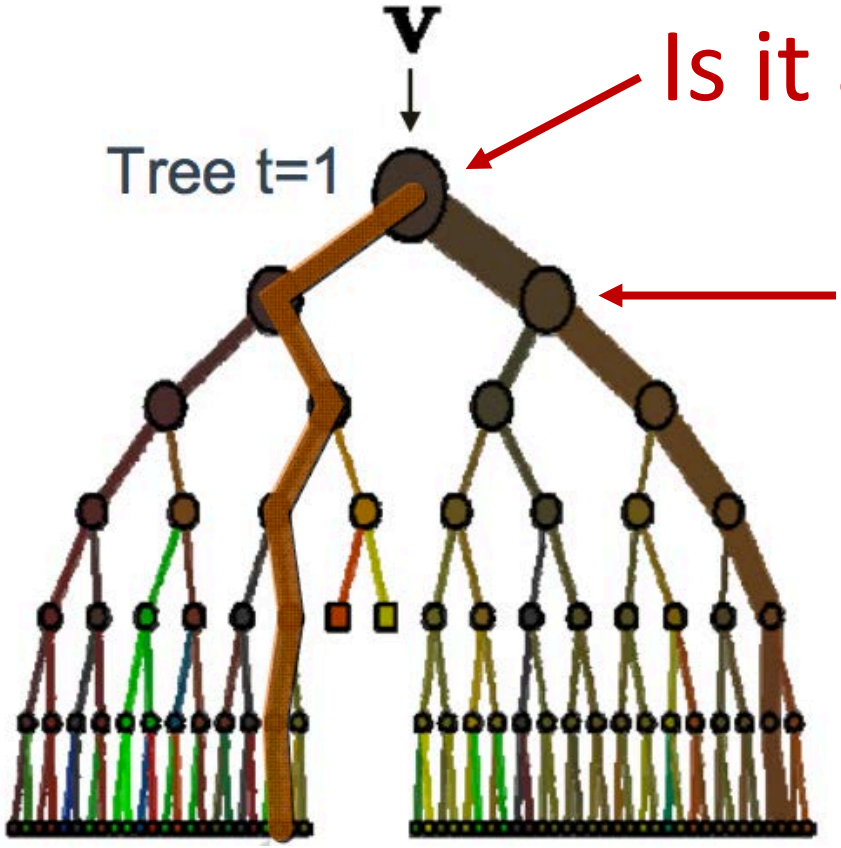
- Machine learning: the computer “learns” how the predictor variables contribute the response variable
- Ensemble: this method uses an ensemble (a group of items viewed as a whole rather than individually) of decision trees to obtain better predictive performance

RANDOM FOREST



Is it an action movie?

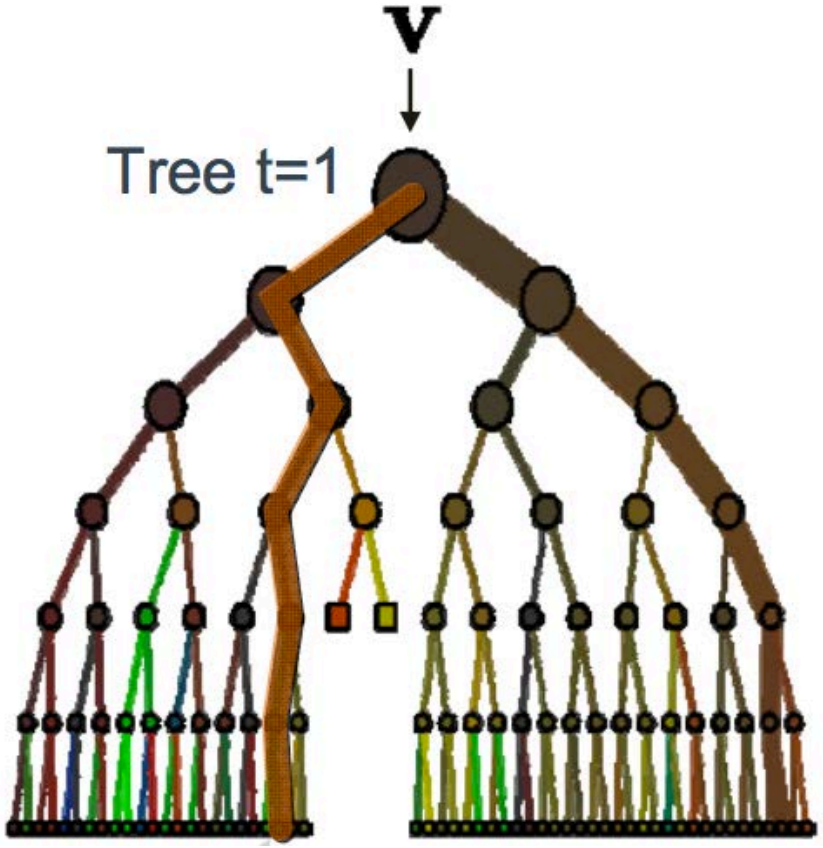
RANDOM FOREST



Is it an action movie?

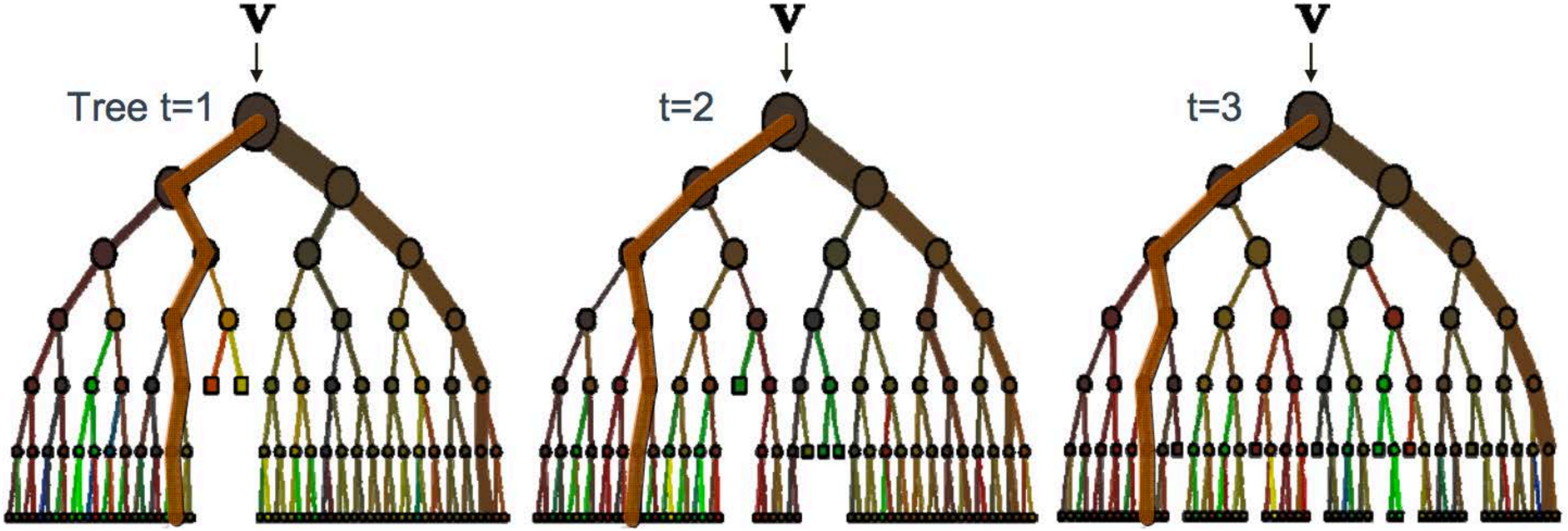
IMDB score >5 or <5?

RANDOM FOREST

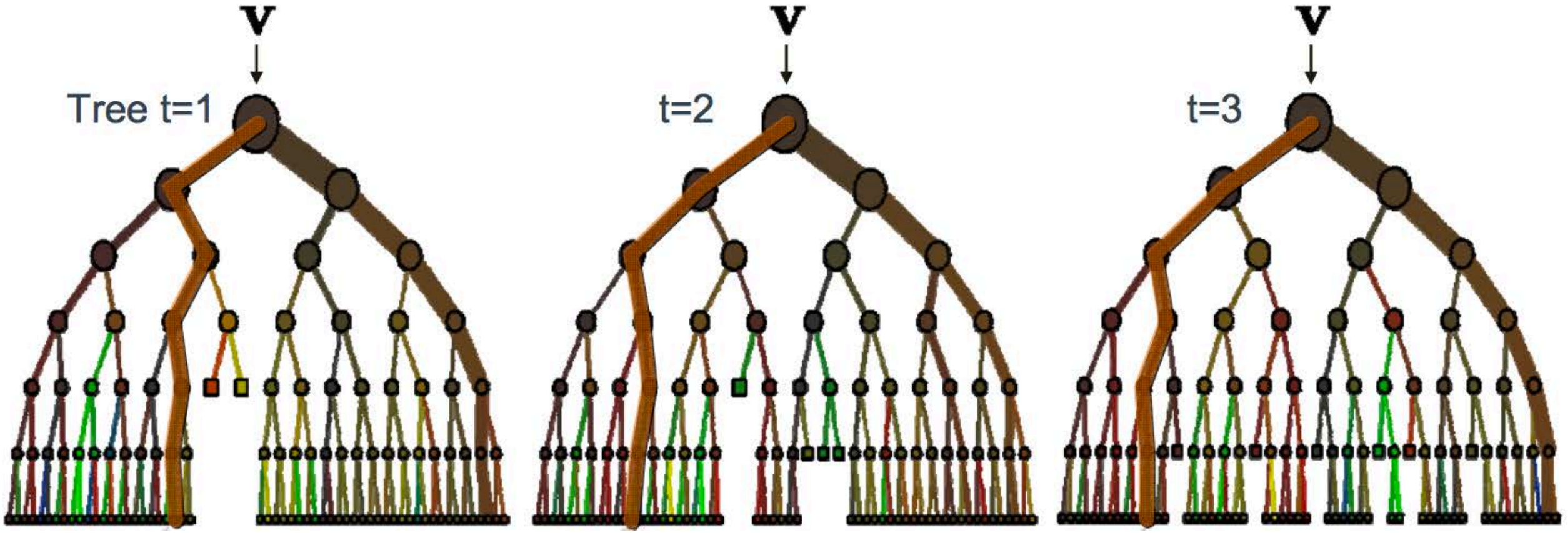


My scores are at the tips of the tree

RANDOM FOREST

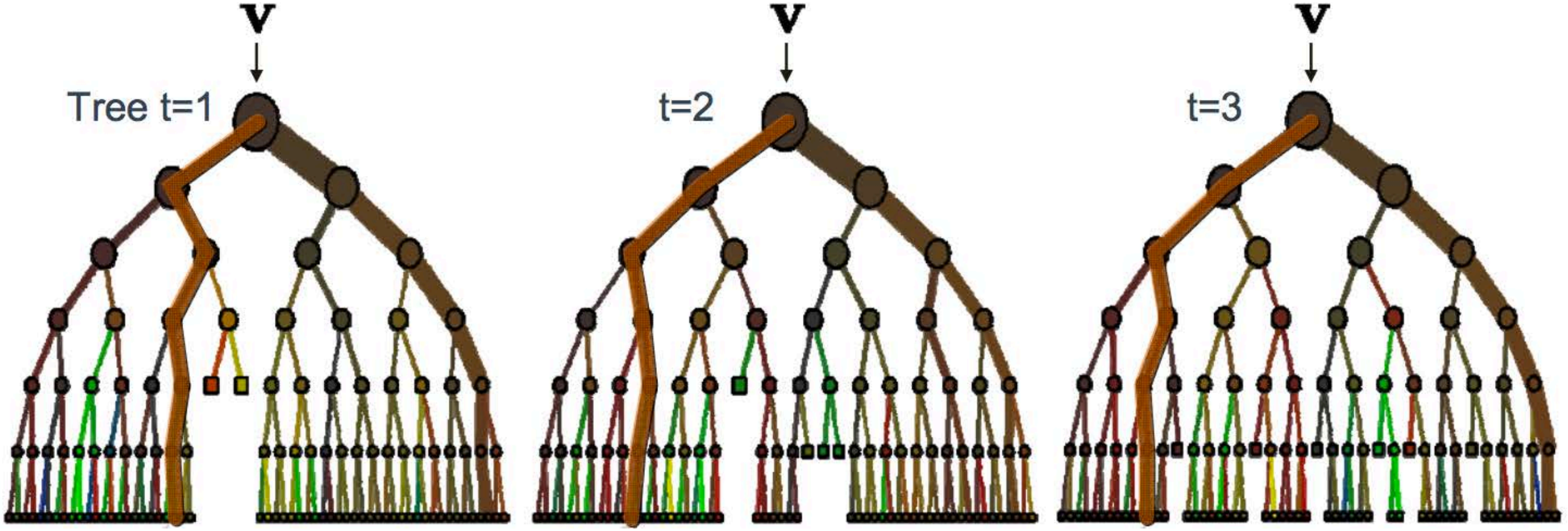


RANDOM FOREST



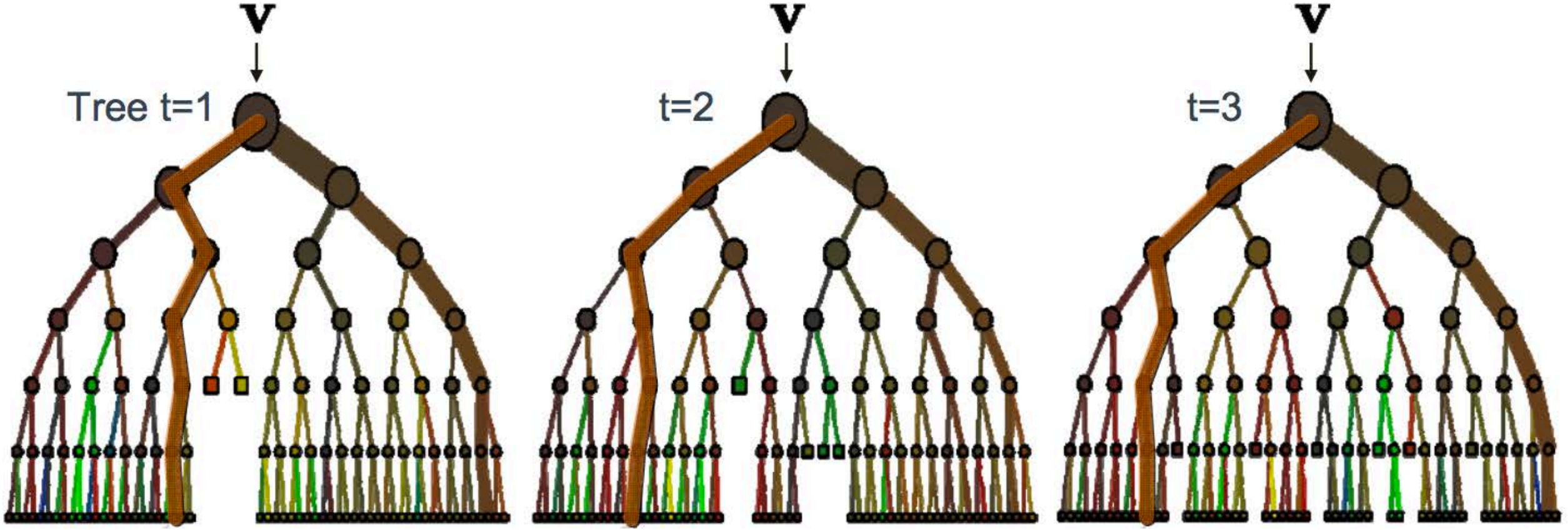
Where does an observation with no known response fall?

RANDOM FOREST



Final answer is consensus of all trees (a vote)

RANDOM FOREST



If continuous, final answer is average of all trees

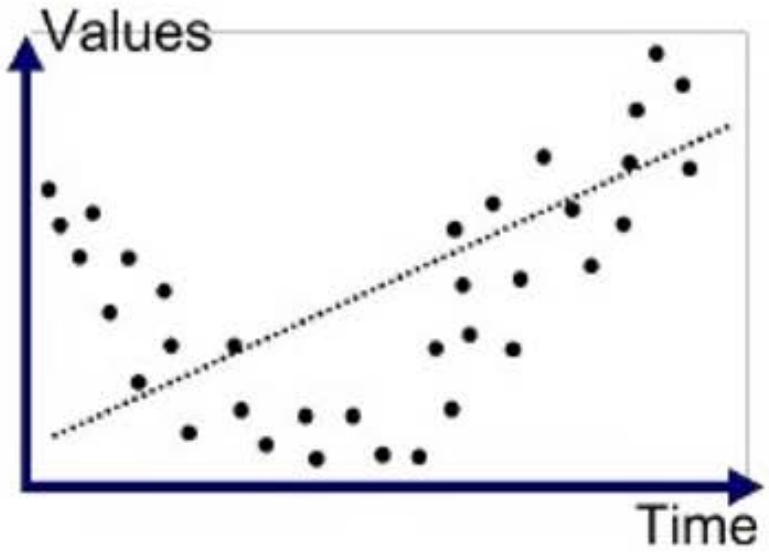
Random Forest

- Bagging (**b**ootstrap **agg**regat**ing**): generate *m* new training sets, of size *n*, by randomly sampling from the full data set with replacement (70/30)
- Random subspace: each decision tree consists of one of these samples from the full dataset

Random Forest

- Bagging (**b**ootstrap **agg**regat**ing**):
  - Reduces variance
  - Avoids overfitting
  - Similar to model averaging

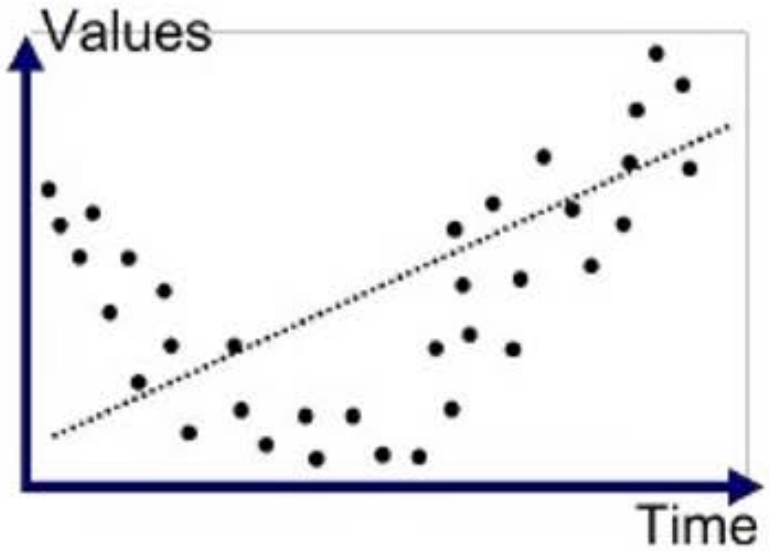
Model fit



Underfitted

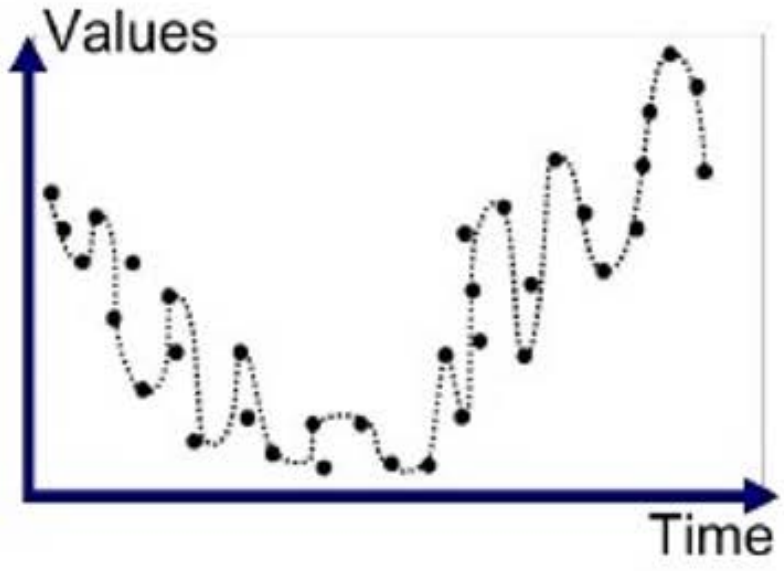
High bias

Model fit



Underfitted

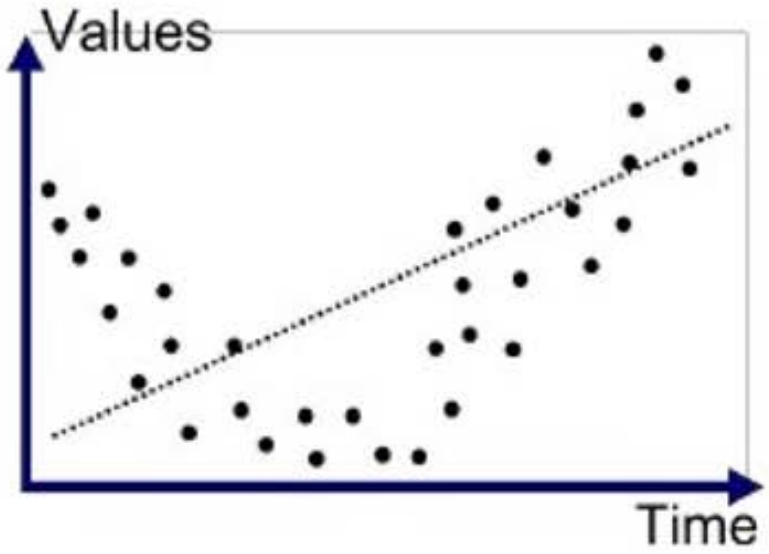
High bias



Overfitted

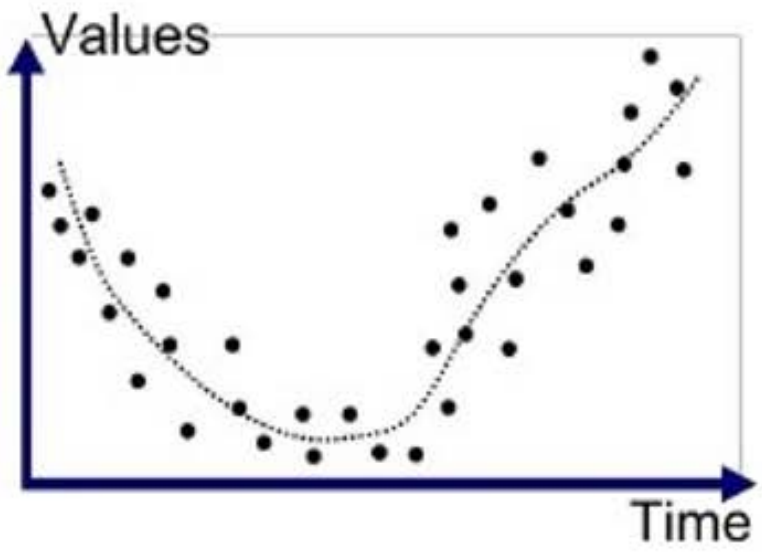
High variance

Model fit

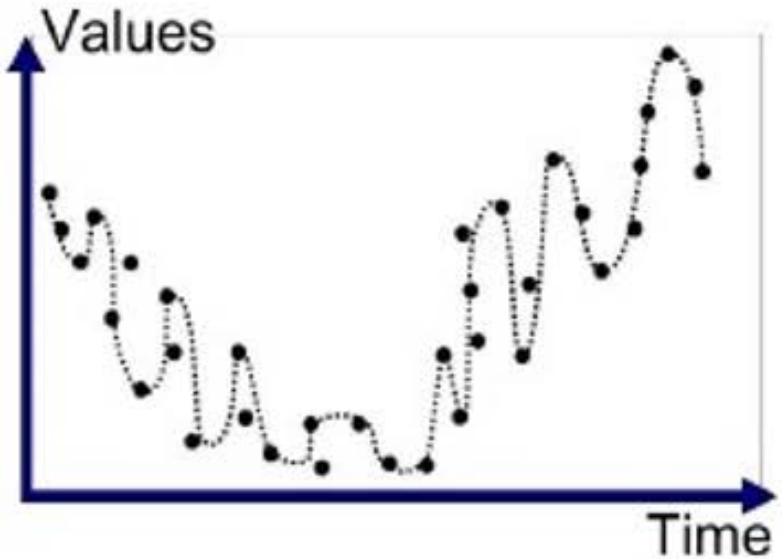


Underfitted

High bias



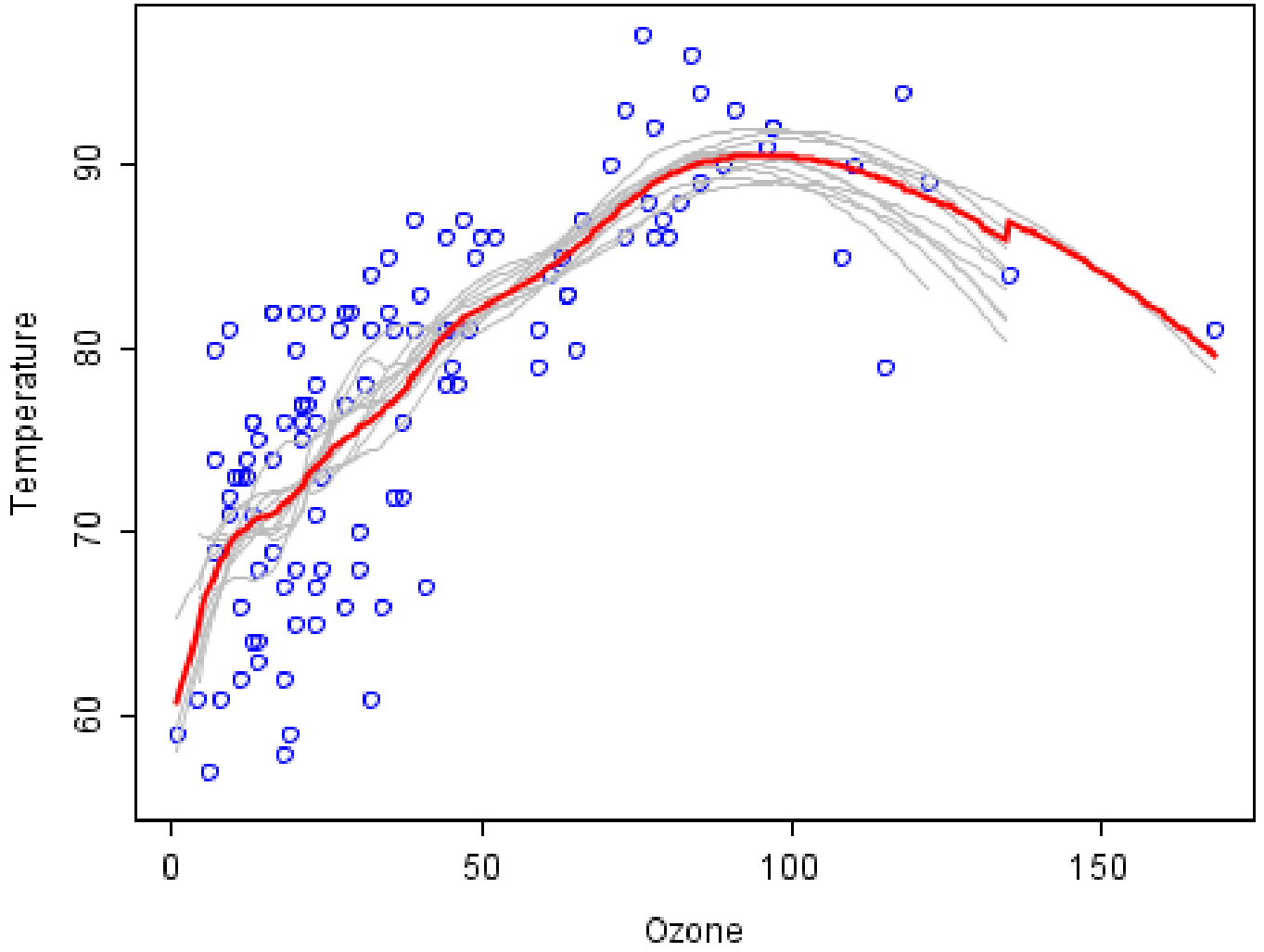
Good Fit/Robust



Overfitted

High variance

Bagging



(Rousseeuw & Leroy 1986)

Identifying cryptic diversity with predictive phylogeography

Anahí Espíndola, Megan Ruffley, Megan L. Smith,* Bryan C. Carstens, David C. Tank, & Jack Sullivan

PROCEEDINGS B

rspb.royalsocietypublishing.org

Identifying cryptic diversity with predictive phylogeography

Anahí Espíndola[1,2], Megan Ruffley[1,2], Megan L. Smith[3], Bryan C. Carstens[3], David C. Tank[1,2] and Jack Sullivan[1,2]

cryptic diversity: the presence of deep genetic divergence within a nominal species with no accompanying fixed morphological differences

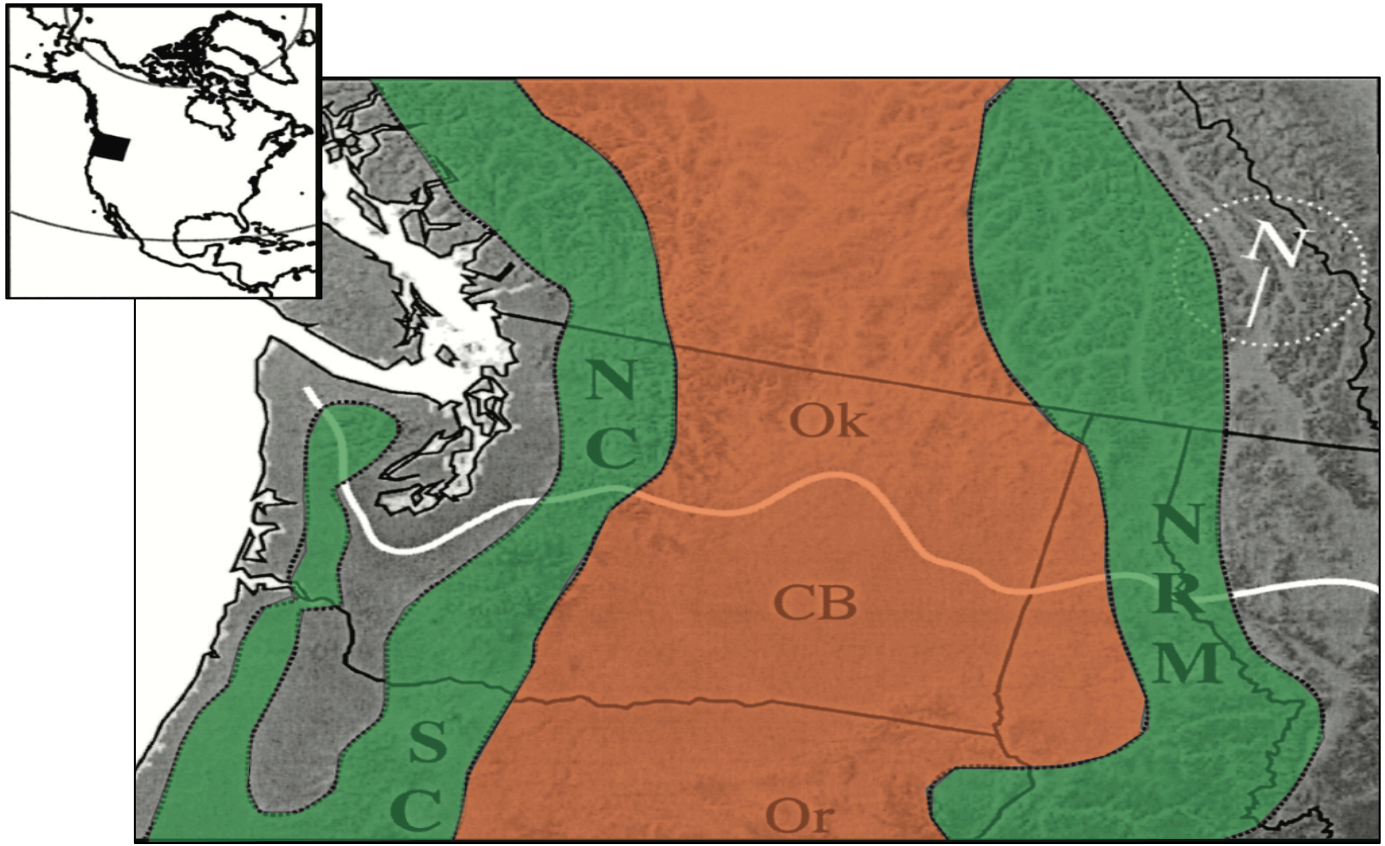
Cryptic diversity

- estimates of species richness and endemism
- definitions of evolutionarily significant units
- understanding of important ecological and evolutionary processes

An approach for the rapid discovery of cryptic diversity…

1) Collect occurrence, environmental, and genetic data from codistributed taxa.
2) Classify reference taxa as cryptic or non-cryptic.
3) Build a Random Forest (RF) classifier using this reference data.
4) Assess the error rate of the classifier.
5) Apply classification function to predict presence or absence of cryptic diversity in new taxa.

Our Focal System: The Pacific Northwest



Step 1: Collect occurrence, taxonomic, and genetic data.

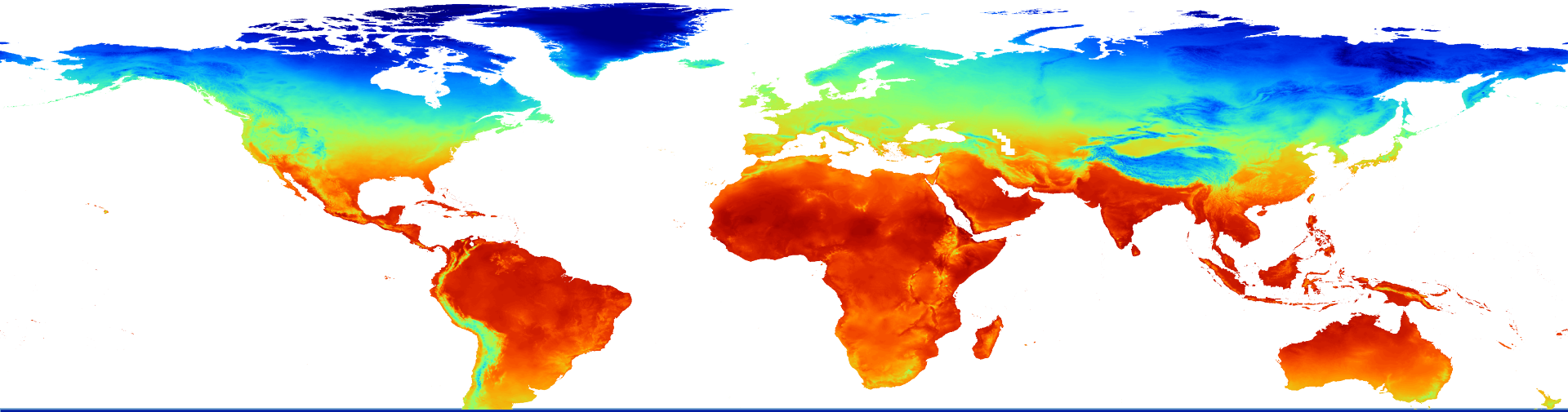
7 taxa

> 1900 localities

> 700 sequences

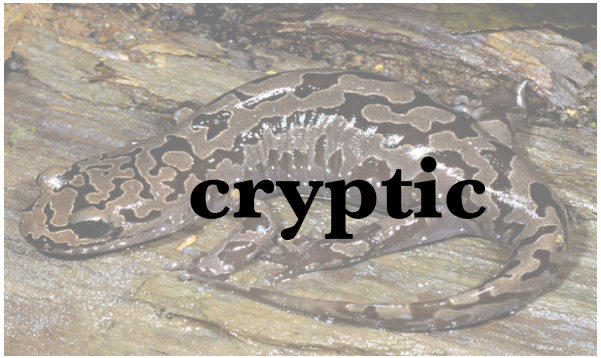
GBIF
Global Biodiversity Information Facility

NCBI



Step 2: Classify reference taxa as cryptic or non-cryptic.

using common approaches for phylogeography



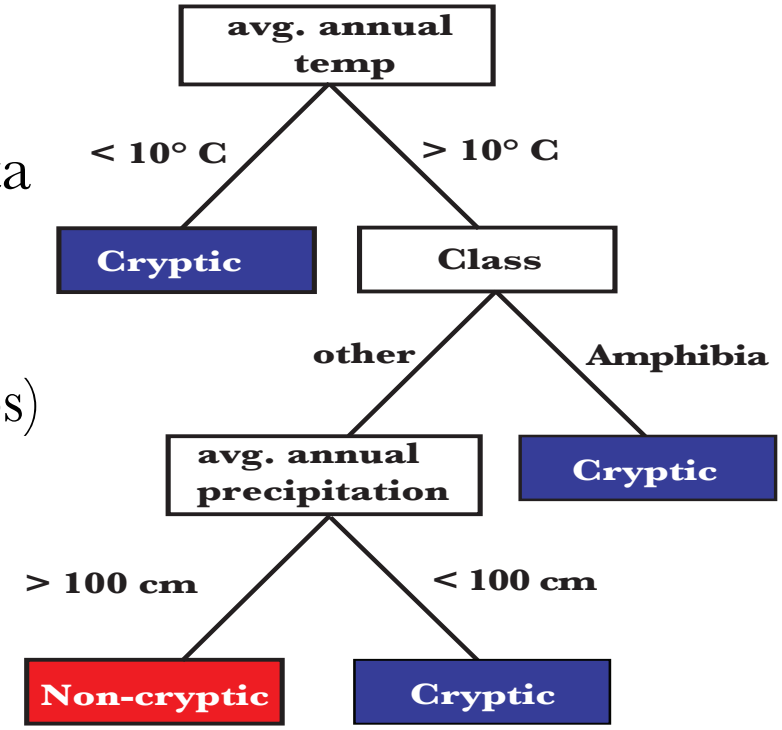
Step 3: Build a Random Forest Classifier.

Decision trees:

Each node is a dichotomization of the data based on predictors.

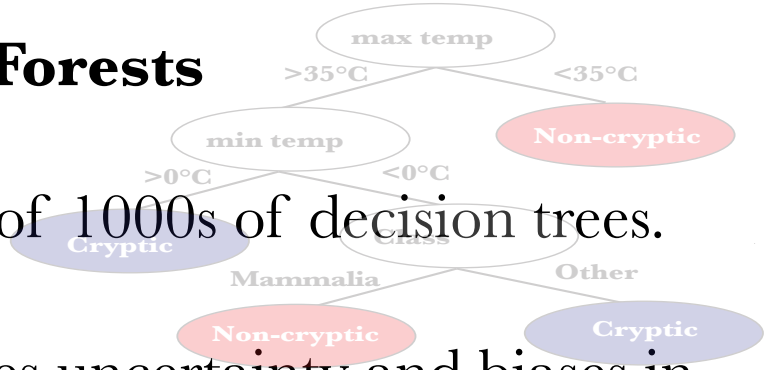
Observations are classified at leaves (or tips) of the trees.

Can be used to classify new data.



Step 3: Build a Random Forest Classifier

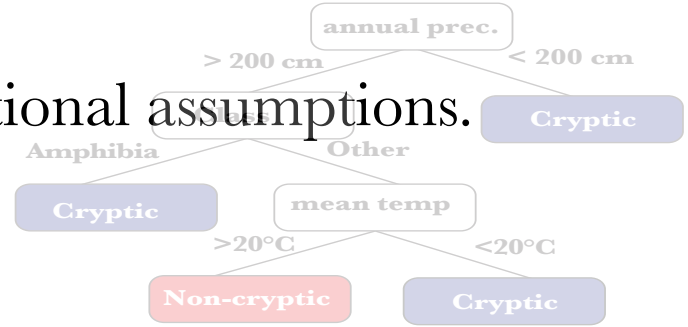
Random Forests



Collection of 1000s of decision trees.

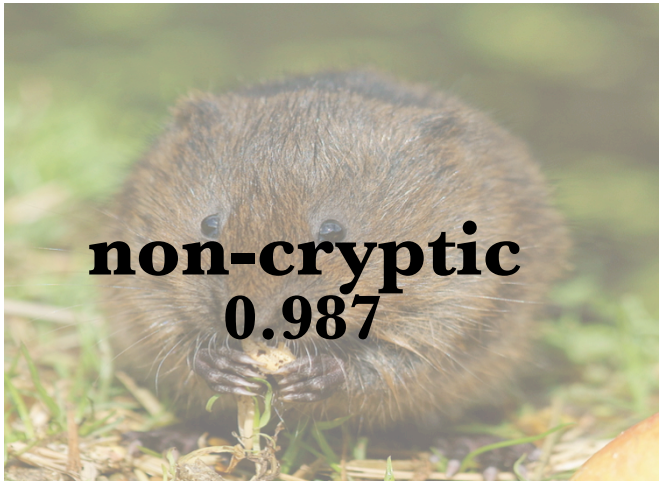
Incorporates uncertainty and biases in classification.

No distributional assumptions.



Step 4: Assess Error Rates.

Jackknife approach: train classification forest on all taxa but one, and apply classification forest to omitted taxon.



overall accuracy: 98.79 %

Step 5: Apply RF Classifier to unstudied taxa.

Haplotrema vancouverense

Alnus rubra

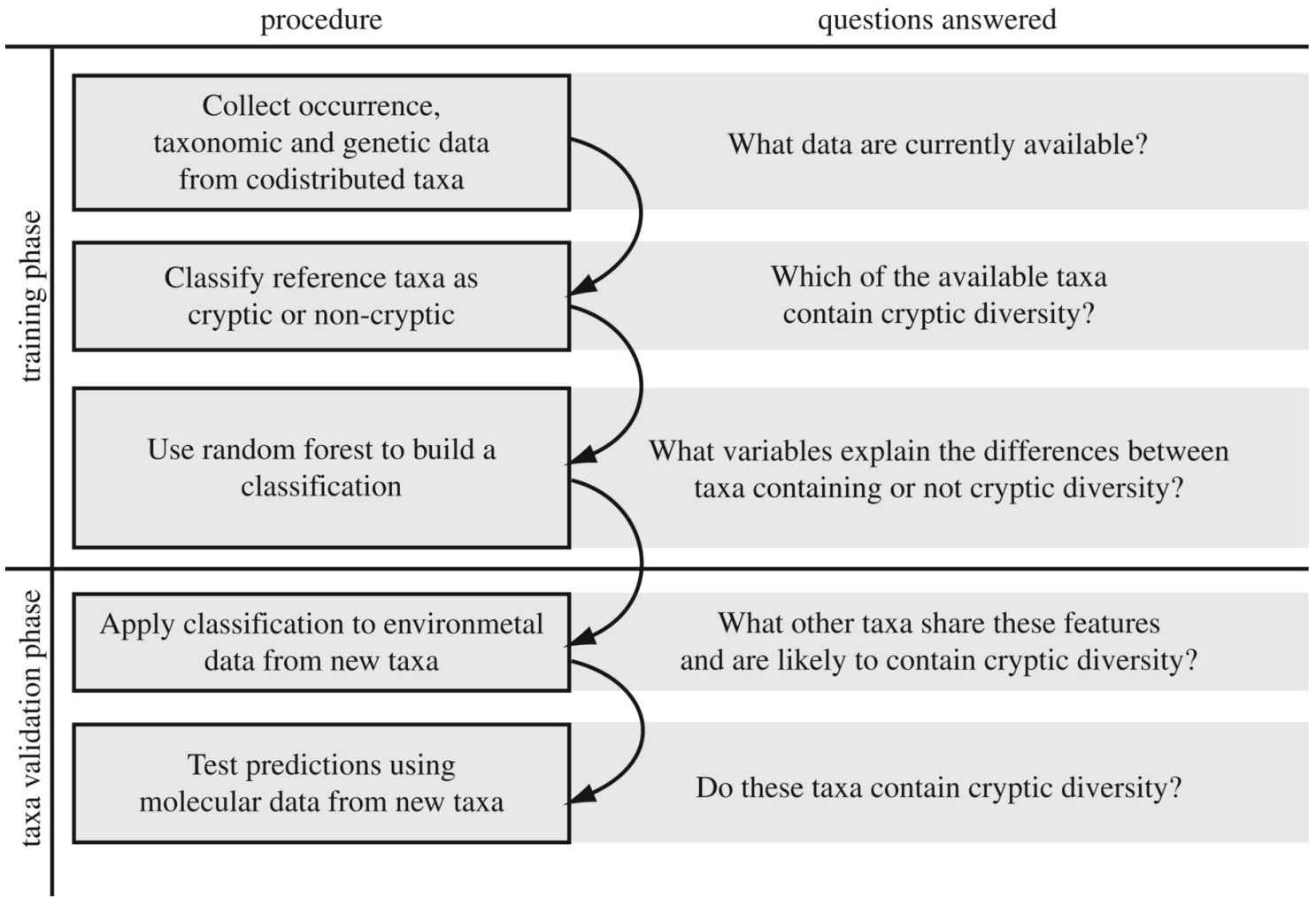
Prediction	Results
Non-cryptic (0.982)	
Non-cryptic (0.981)	

Step 5: Apply RF Classifier to unstudied taxa.

Haplotrema vancouverense

Alnus rubra

Prediction	Results
Non-cryptic (0.982)	Non-cryptic
Non-cryptic (0.981)	Non-cryptic



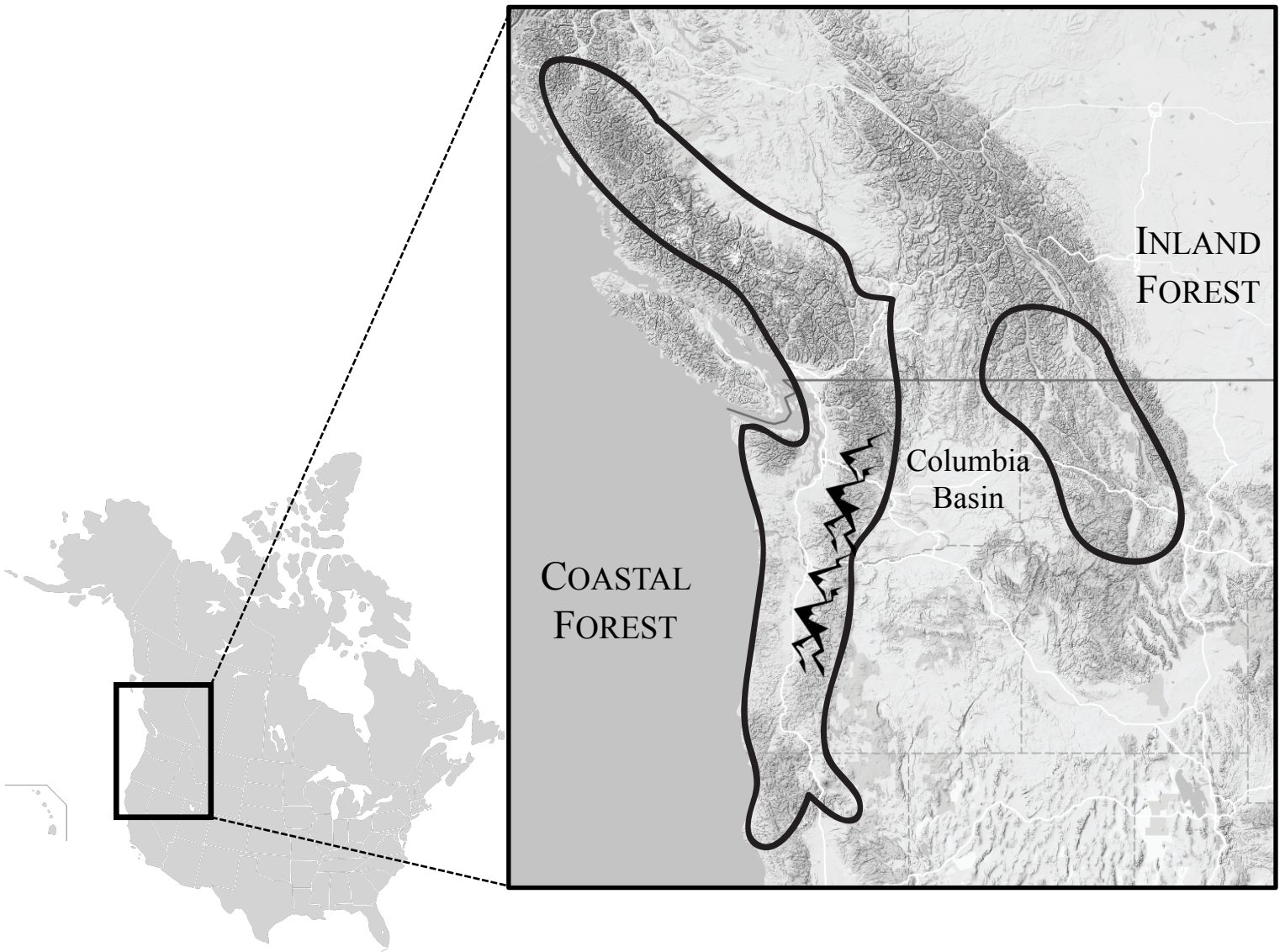
Anahí Espíndola et al. Proc. R. Soc. B 2016;283:20161529

PROCEEDINGS OF THE ROYAL SOCIETY B

Conclusions and Future Directions

- Predictive phylogeography allows for the rapid identification of cryptic diversity.
- Even with limited ecological data (climatic and taxonomic data only) we can make accurate predictions about the presence or absence of cryptic diversity.
- Future work will incorporate more ecological data and a larger reference set of taxa to improve our classification accuracy (You guys get to try this, later!!)

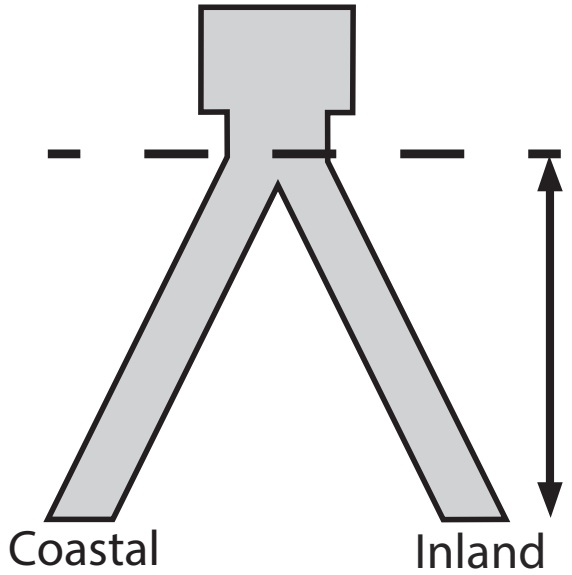
Phylogeographic Model Selection

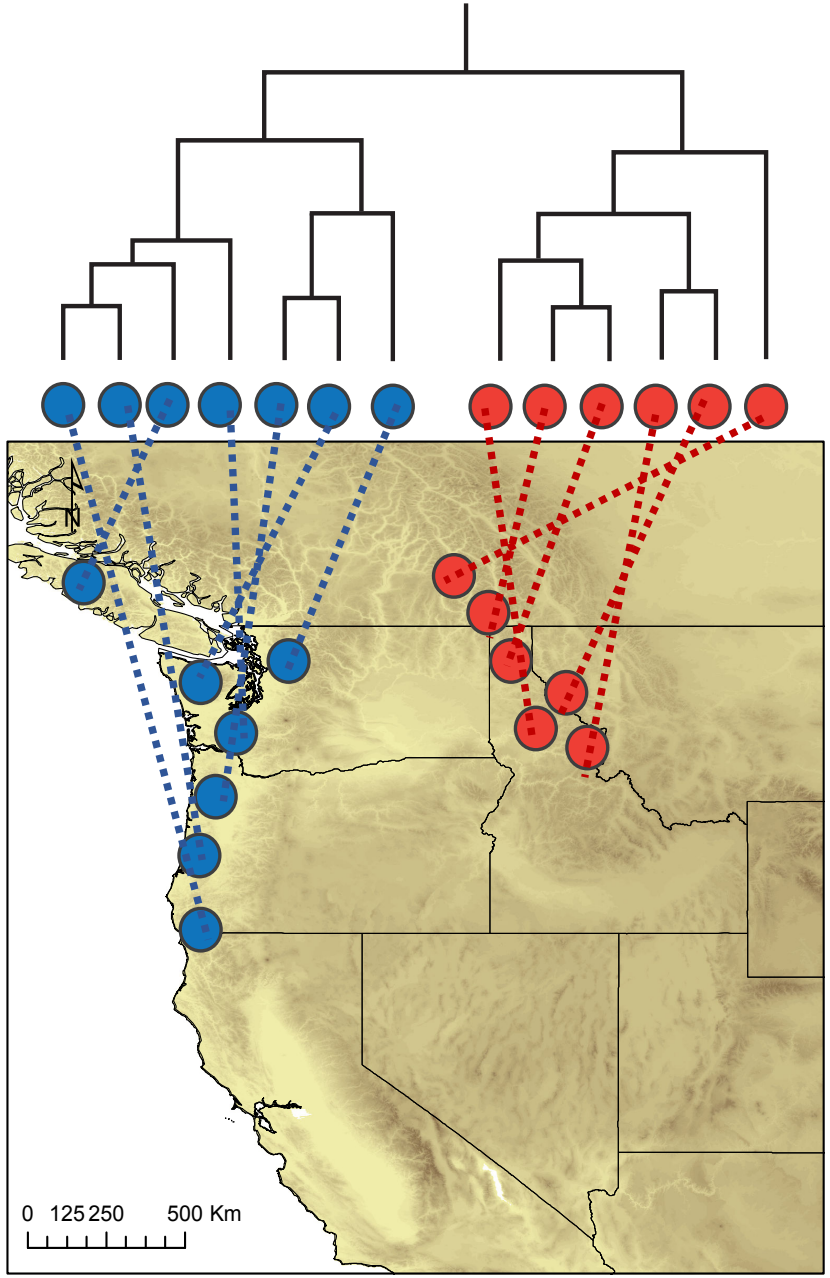


Phylogeography

…uses phylogenetics to explain how the genetic variation of a species is geographically distributed

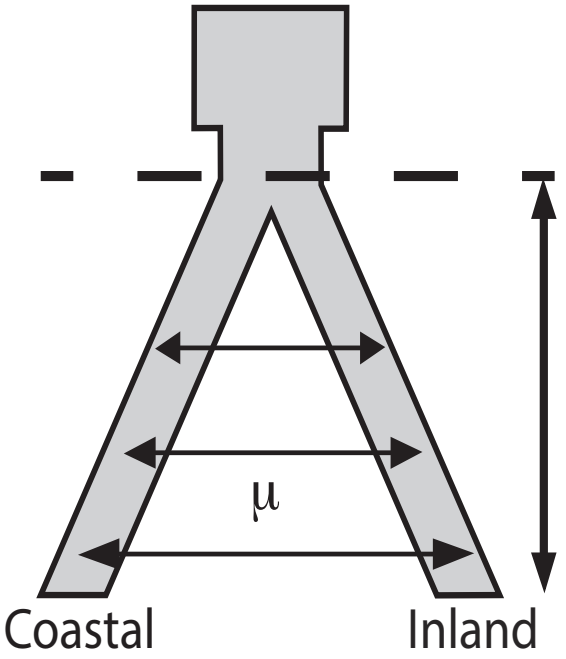
Ancient Vicariance

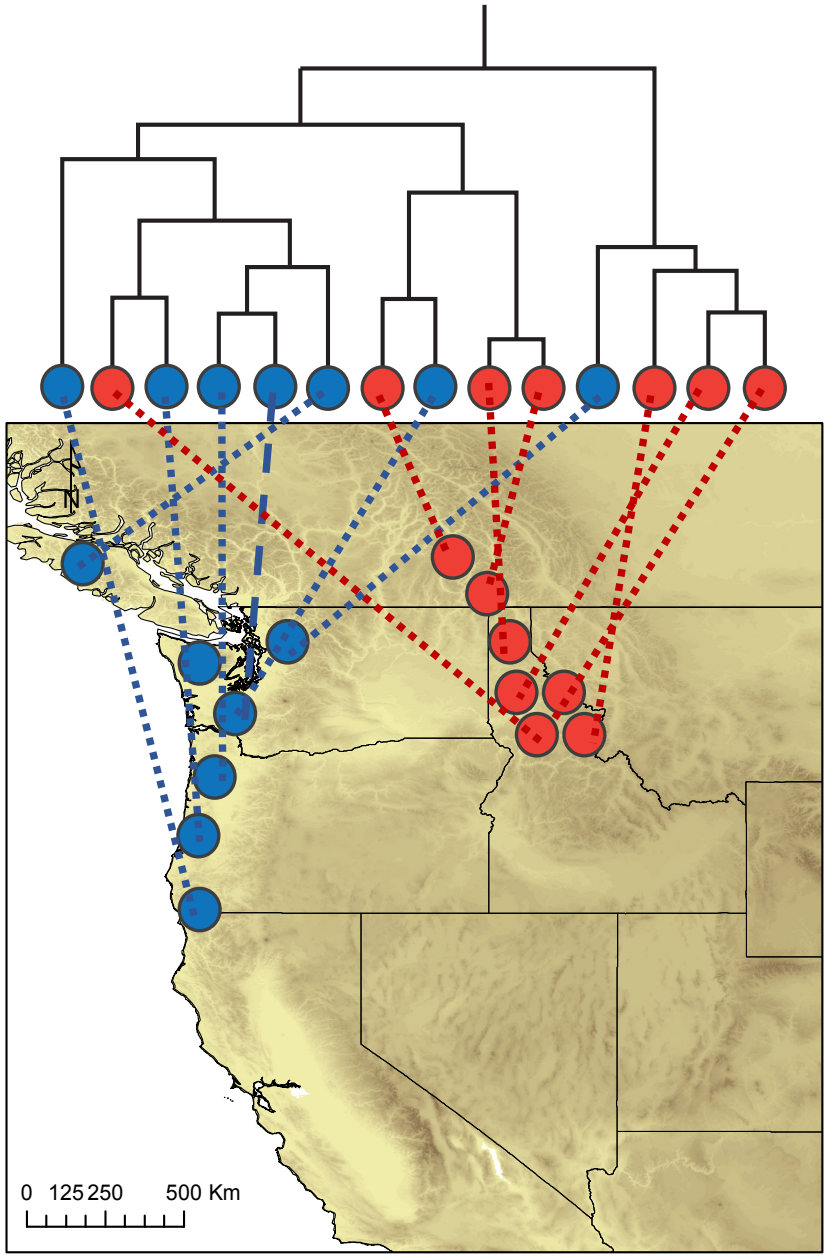




Phylogeography

AV with Migration

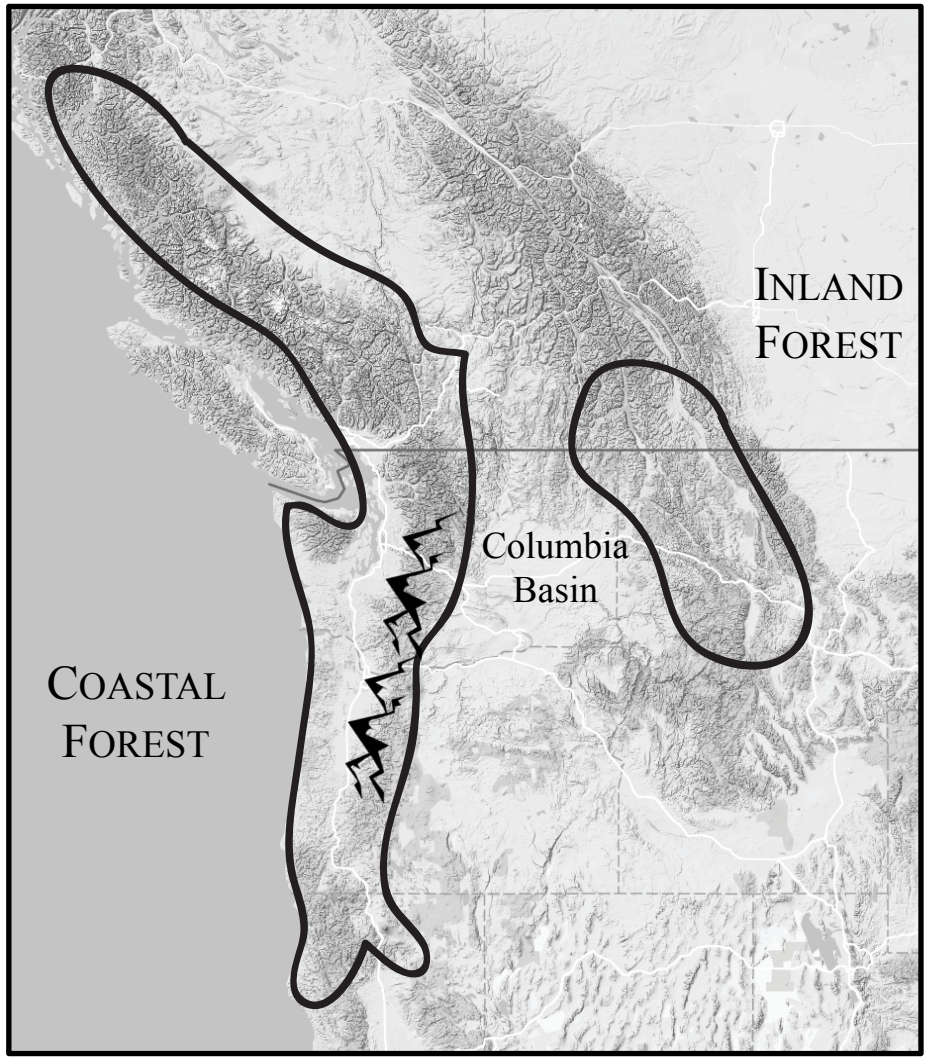




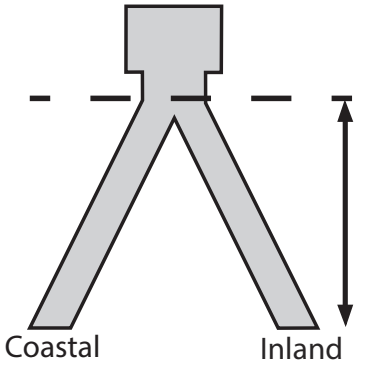
Phylogeographic Model Selection

Microtus richardsoni

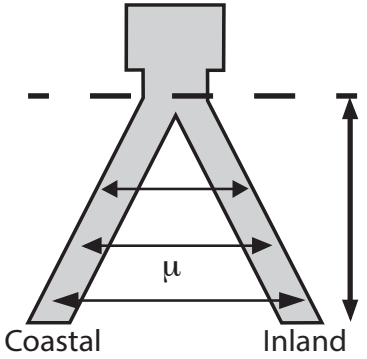
Sampling:
Inland = 34
Coastal = 24



Ancient Vicariance



AV with Migration

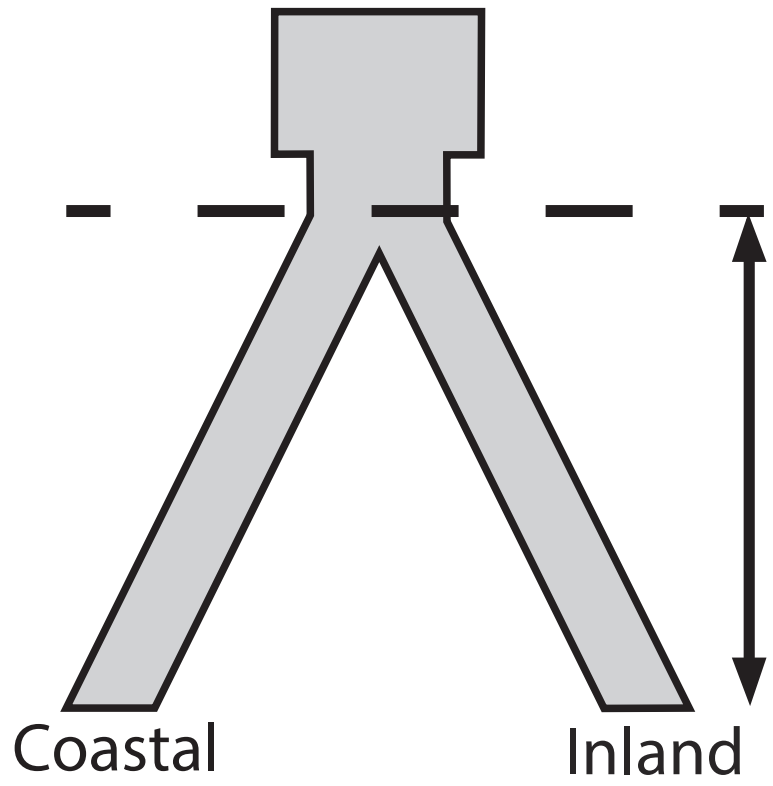


Model Selection with Random Forest

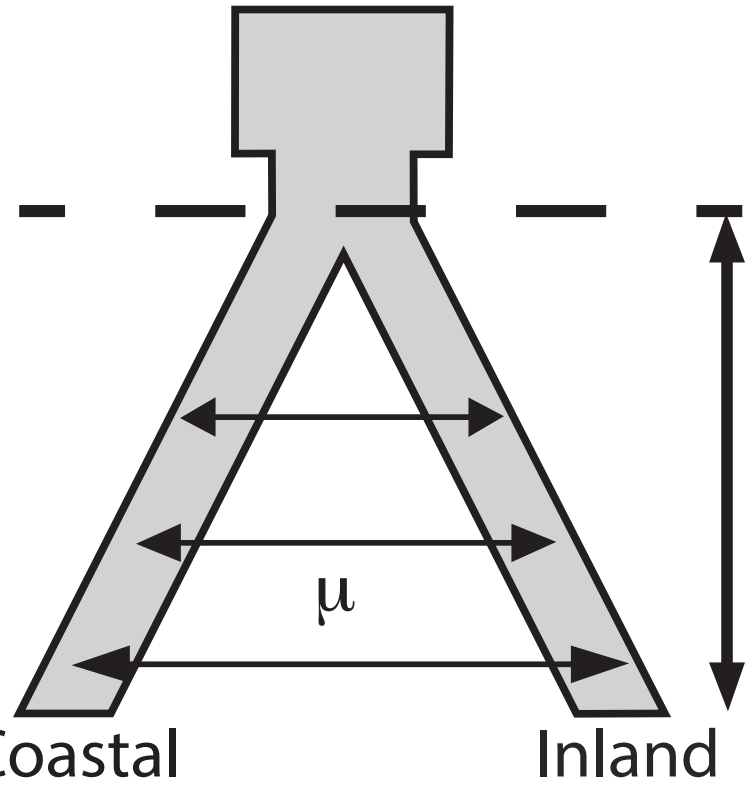
- Approximate method for model selection
- We will need to simulate data under each phylogeographic hypothesis (model : Ancient Vicariance, model2: AV with migration)
- Summarizing that data in the form of summary statistics
- Use those summary statistics as predictor variables to construct a random forest classifier with model identities as the response variable

Simulate Data

Ancient Vicariance



AV with Migration



Summarize Data

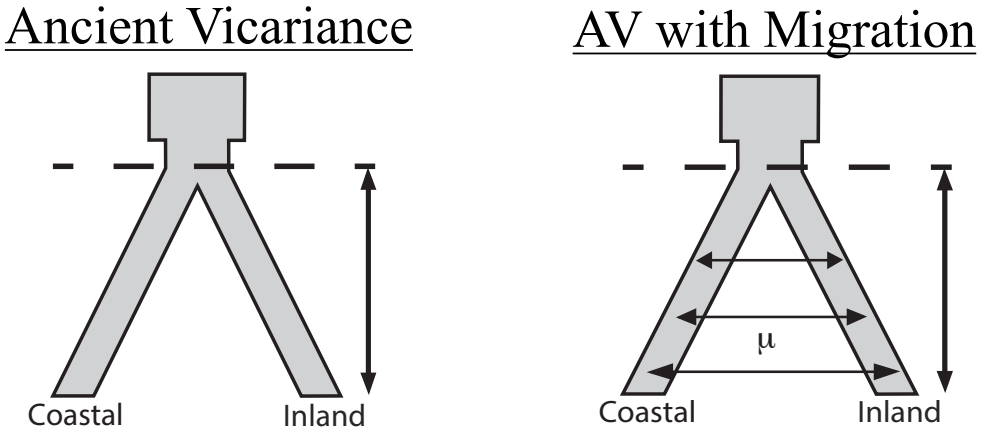
Predictor variables

Response variable

sim.num	pi	ss	D	thetaH	H	pi.w.1	pi.w.2	pi.btw.12	mod
1	5.580762	29	-0.354638	2.735027	2.845735	5.6969697	5.2862319	5.600490	mod2
2	3.792498	22	-0.637227	2.207502	1.584997	4.5614973	2.2173913	3.796569	mod2
3	113.176649	293	2.810211	106.156685	7.019964	15.5739750	13.1884058	214.098039	mod1
4	3.427707	14	0.394909	2.537205	0.890502	3.3475936	3.5869565	3.428922	mod2
5	33.879613	77	3.585284	35.488808	-1.609195	1.4242424	2.1594203	66.921569	mod1
6	34.542045	77	3.723002	34.931639	-0.389595	0.6969697	1.4384058	69.007353	mod1
7	8.987296	24	2.336443	7.995160	0.992136	1.0730838	1.4420290	16.980392	mod1
8	141.016334	334	3.408150	137.931034	3.085299	7.0071301	15.7210145	275.526961	mod1
9	3.019964	22	-1.149948	5.471264	-2.451301	3.0713012	2.9311594	3.014706	mod2
10	6.396854	23	0.911333	4.831216	1.565638	5.1925134	7.6920290	6.786765	mod2
11	10.857834	38	1.072249	5.773745	5.084090	11.2352941	9.8876812	10.926471	mod2
12	12.480339	44	1.051015	9.274047	3.206292	12.6185383	12.1630435	12.492647	mod2
13	129.597701	280	4.072221	133.314580	-3.716878	1.9197861	2.0869565	260.504902	mod1
14	14.151240	79	-0.591292	13.918935	0.232305	13.5561497	13.6630435	14.725490	mod2
15	6.904416	29	0.331574	4.569268	2.335148	7.1622103	6.6702899	6.806373	mod2

…100,000 simulations total

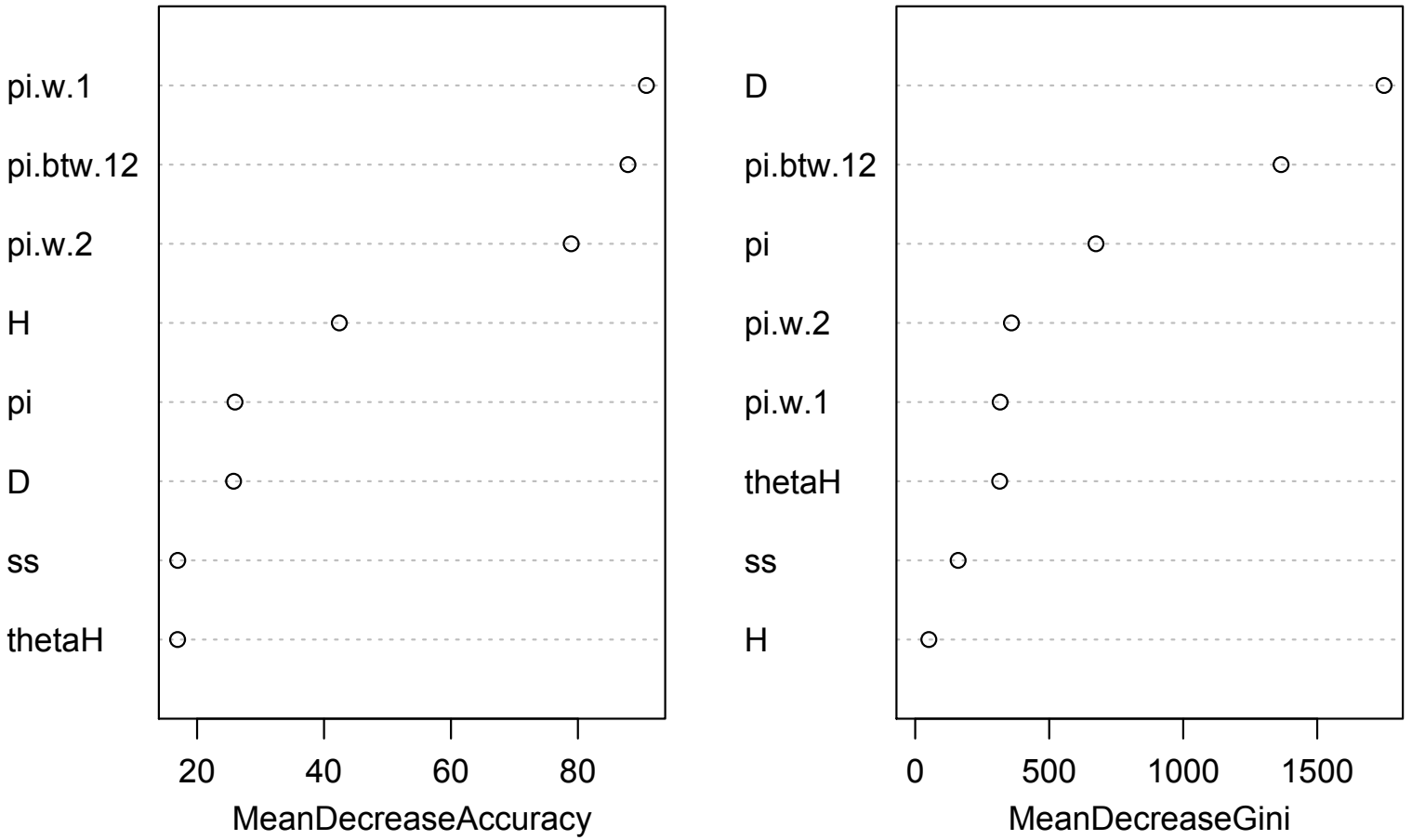
Construct a randomForest classifier




```
Call:
 randomForest(formula = mod ~ ., data = pred.data, ntree = 1000,
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.68%
Confusion matrix:
     mod1 mod2 class.error
mod1 5063   51 0.009972624
mod2   17 4869 0.003479329
```

Look at variable importance



Predict unknown data

Calculate summary statistics for observed data:

Microtus richardsoni cytochrome b gene, partial cds; mitochondrial.

PopSet: 66476210

GenBank FASTA

Go to:

Study Details

Investigating the evolutionary history of the Pacific Northwest mesic forest ecosystem: hypothesis testing within a comparative phylogeographic framework.

Carstens,B.C., Brunsfeld,S.J., Demboski,J.R., Good,J.M. and Sullivan,J.

(2005) Evolution 59:(8)1639-1652

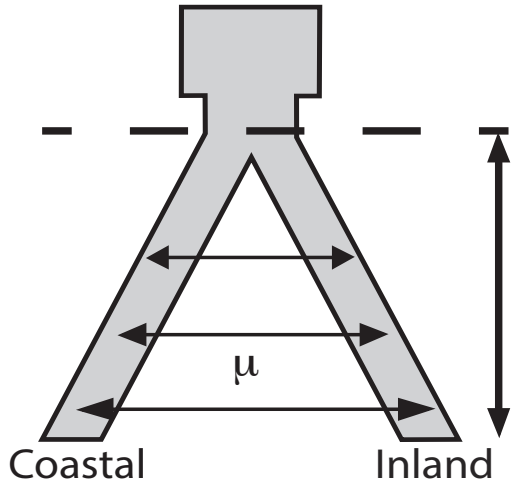
PMID: 16331838 Citation

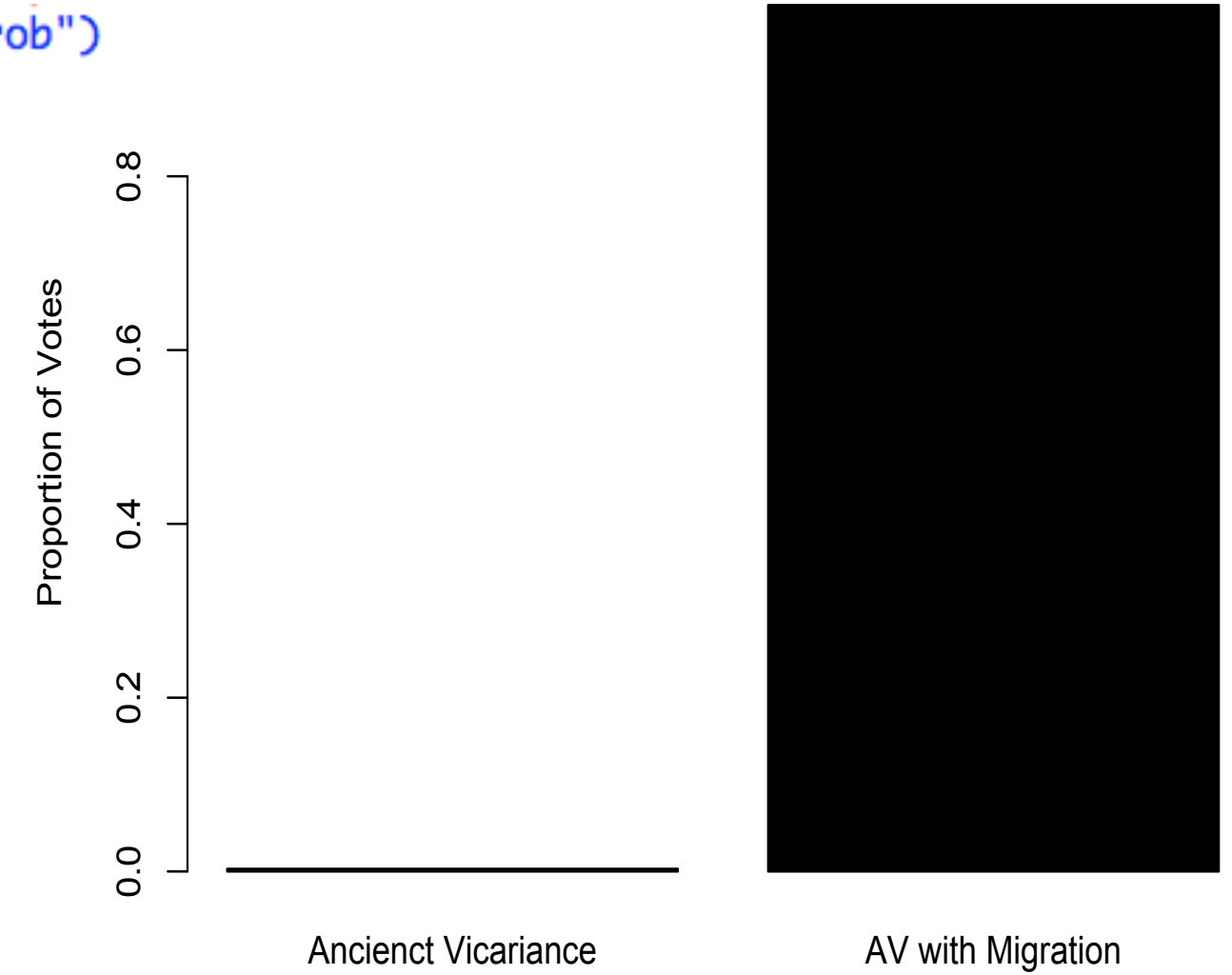
	pi	ss	D	thetaH	H	pi.w.1	pi.w.2	pi.btw.12
ss1	12.06405	59	-0.43723	17.46155	5.3975	10.24137	10.58499	13.681

Predict unknown data

```
> predict(rf, mrich.ss, type="prob")
     mod1  mod2
ss1 0.003 0.997
ss2 0.003 0.997
attr(,"class")
[1] "matrix" "votes"
```

AV with Migration





Predict unknown data

- Compare randomForest predictions to ABC results from Espindola *et al*. 2016

Supplementary Table 2

	Migration 1 (West to East)	Migration 2 (East to West)	Migration 3 (Both directions)	**Best Migration Model**	**Ancient Vicariance**
PNW					
A. montanus - A. truei	0.128	0.115	0.757	0.001	0.999
C. armata	0.457	0.135	0.409	0.865	0.135
D. aterrimus - D. tenebrosus	0.538	0.285	0.177	0.015	0.985
M. richardsoni	0.271	0.151	0.578	**0.999**	**0.001**
P. coeruleum	0.121	0.028	0.851	0.986	0.014
P. vandykei - P. idahoensis	0.158	0.177	0.665	0.000	1.000
S. melanopsis	0.294	0.273	0.434	0.999	0.001

Other examples of Model Selection using *randomForest*

- Tomorrow, June 2nd...

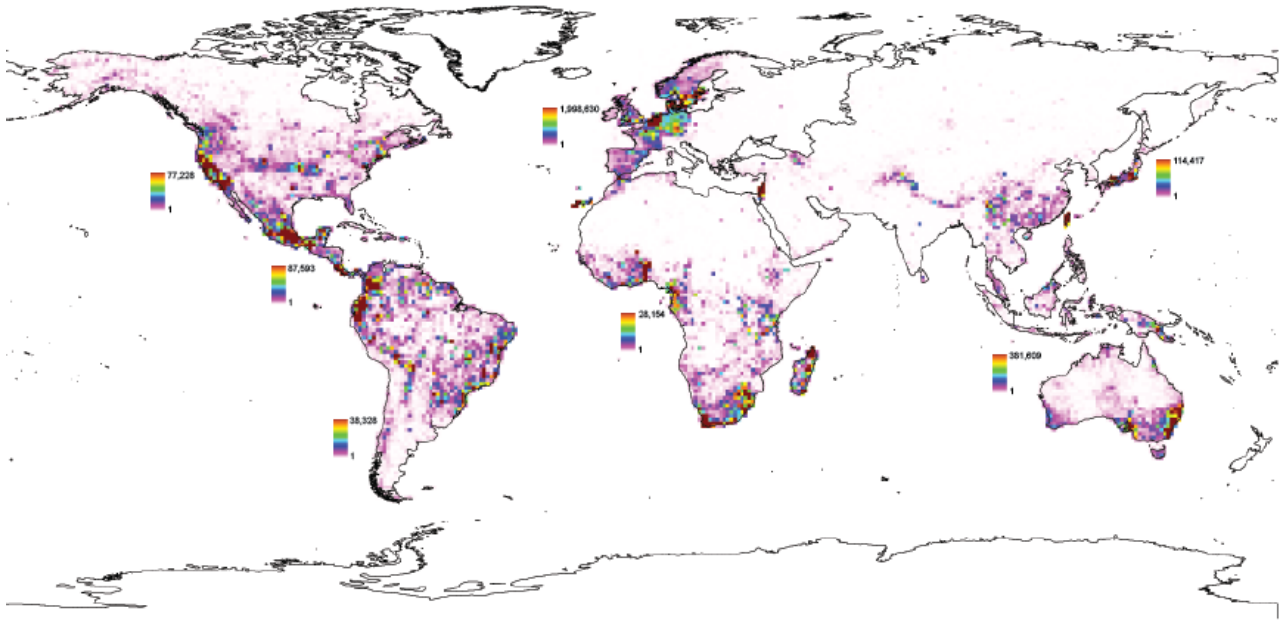
- **3:35 PM** Trait-Mediated Community Assembly Models Identified through Machine Learning and Approximate Bayesian Computation
  - Megan Ruffley, Katie Peterson, Bob Week, David C. Tank, Luke Harmon

- **3:40 PM** Species Delimitation Using Random Forests and the Site Frequency Spectrum
  - Megan L. Smith, Bryan C. Carstens

Practice data set

- Use open data to create a classifier that predicts which unlisted taxa are likely to fall in non-Least Concern categories

Used GPS points from GBIF for all land plants to extract environmental and geographic information to use as predictor variables



Pelletier et al. *in review*

Practice data set

TUTORIAL

Found here:

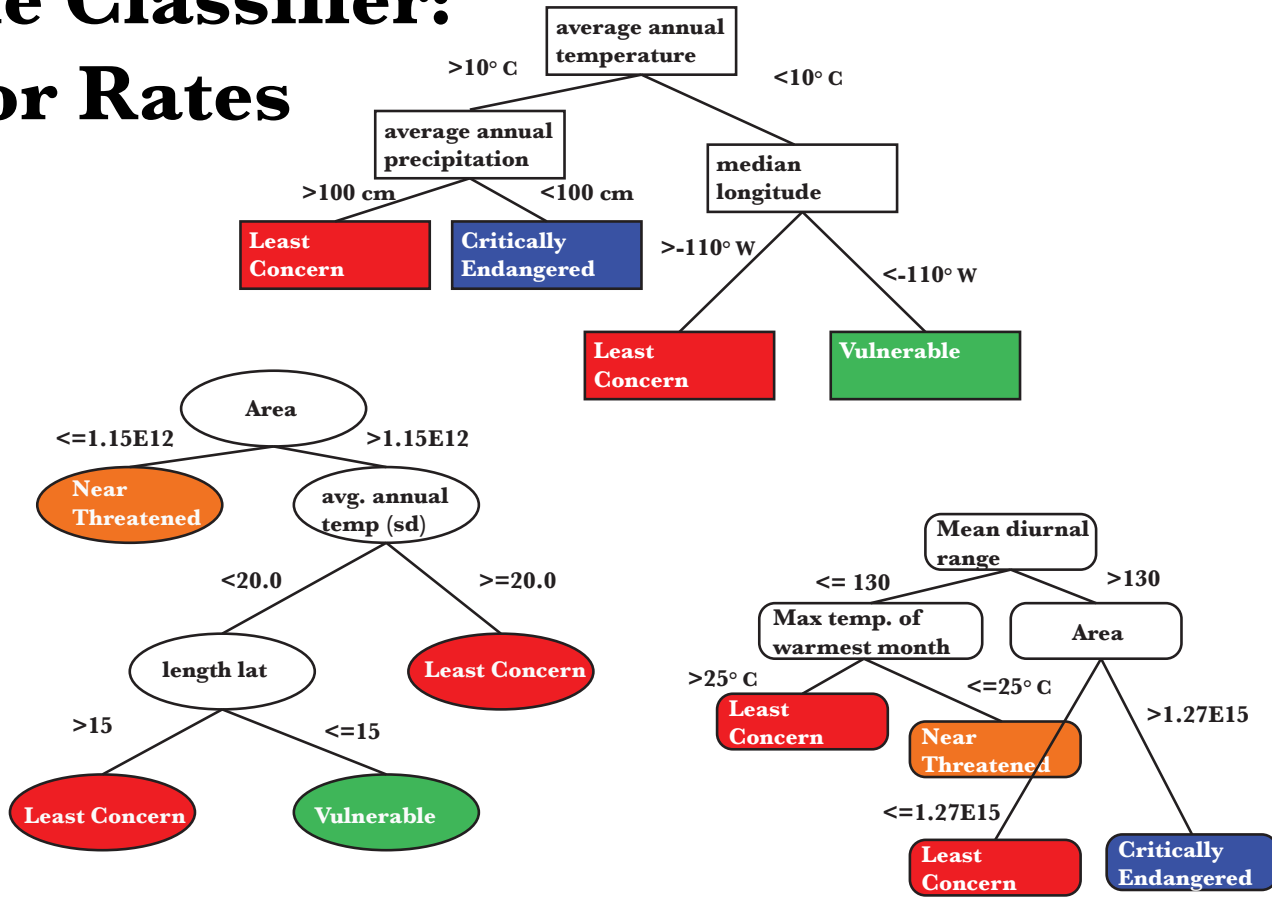
https://predictivephylogeography-ssb2018.github.io/

1. Introduction
2. Explore the data
3. Build the classifier
4. Looking at the forest

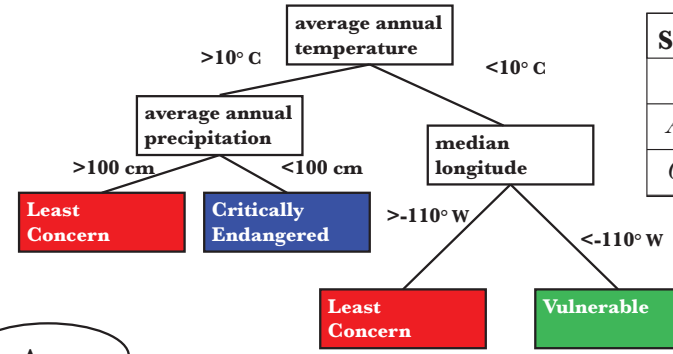
TUTORIAL

5, 6 & 7: Evaluating and Improving the Classifier

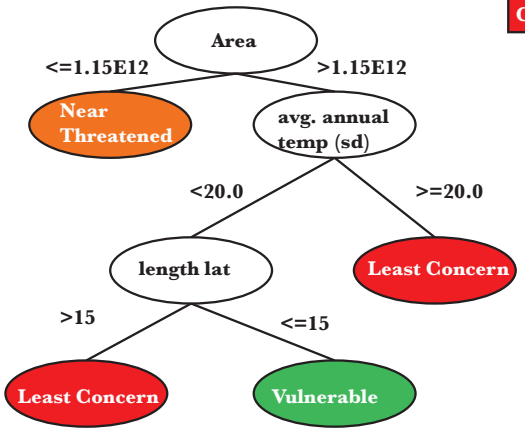
5. Evaluating the Classifier: Out-of-Bag Error Rates



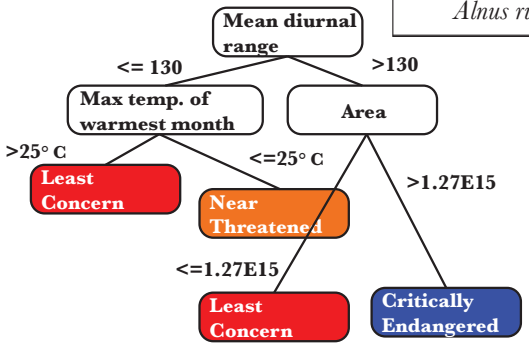
Out-of-Bag Error Rates



Species	n.gps	abs_m
Carex bicolor	127	
Astragalus miser	50	
Opuntia robusta	405	

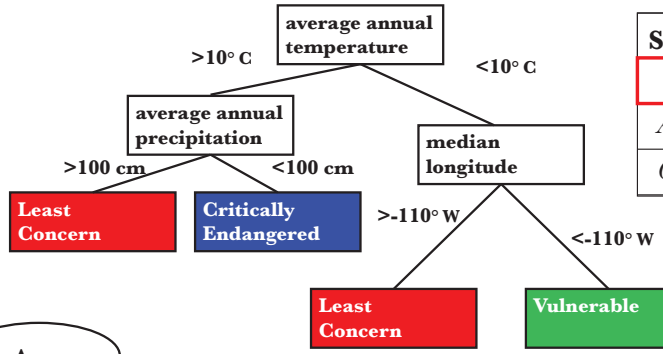


Species	n.gps	abs_m
Geum rivale	193	
Elodea canadensis	67	
Alnus rubra	1046	

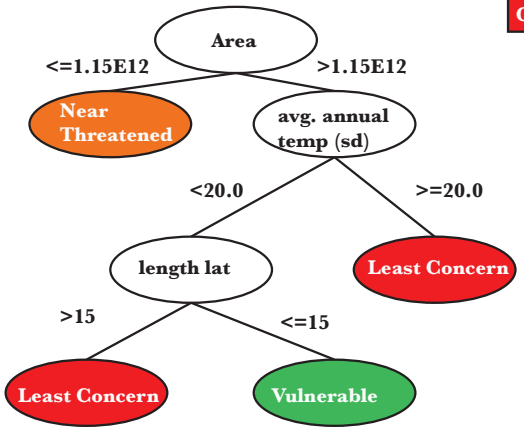


Species	n.gps	abs_m
Panicum strigosum	76	
Pinus ponderosa	975	
Quercus rubra	1574	

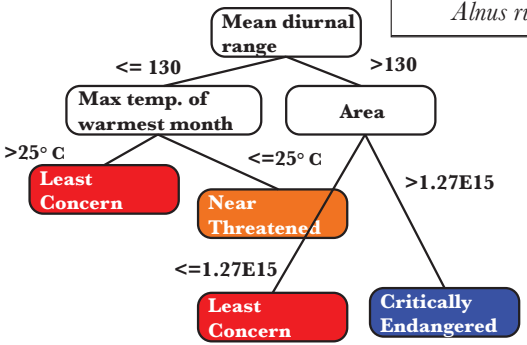
Out-of-Bag Error Rates



Species	n.gps	abs_m
Carex bicolor	127	
Astragalus miser	50	
Opuntia robusta	405	



Species	n.gps	abs_m
Geum rivale	193	
Elodea canadensis	67	
Alnus rubra	1046	



Species	n.gps	abs_m
Panicum strigosum	76	
Pinus ponderosa	975	
Quercus rubra	1574	

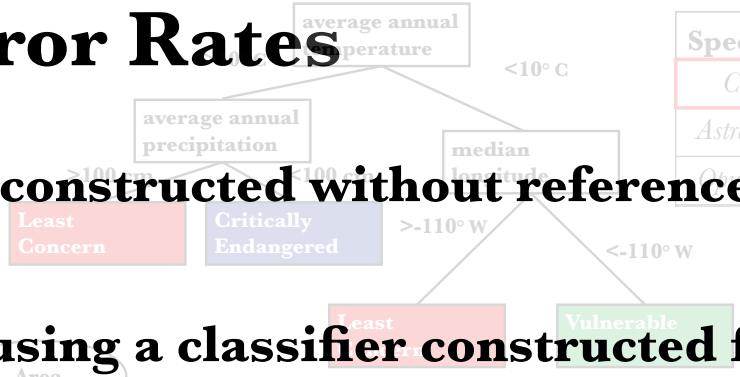
Out-of-Bag Error Rates

1. **Find decision trees constructed without reference to *Carex bicolor*.**

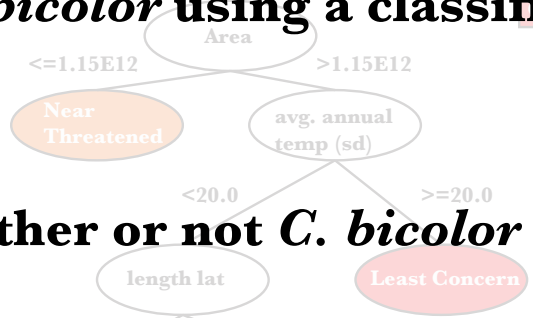
2. **Classify *C. bicolor* using a classifier constructed from those decision trees.**

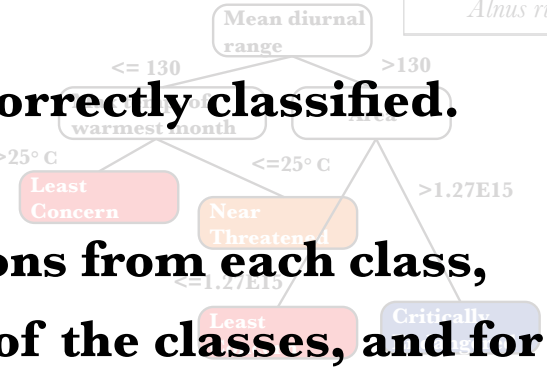
3. **Record whether or not *C. bicolor* is correctly classified.**

4. **When we do this for many observations from each class, we can estimate error rates for each of the classes, and for the classifier as a whole.**



Species	n.gps	abs_m
Carex bicolor	127	
Astragalus miser	50	





Species	n.gps	abs_m
Geum rivale	193	
Elodea canadensis	67	
Alnus rubra	1046	

Species	n.gps	abs_m
Pinus ponderosa	975	
Quercus rubra	1574	

Other Methods of Assessing Error Rates

- **Cross-validation**

  Omit a subset of the data when building the classifier, and then apply the classifier to the omitted data.

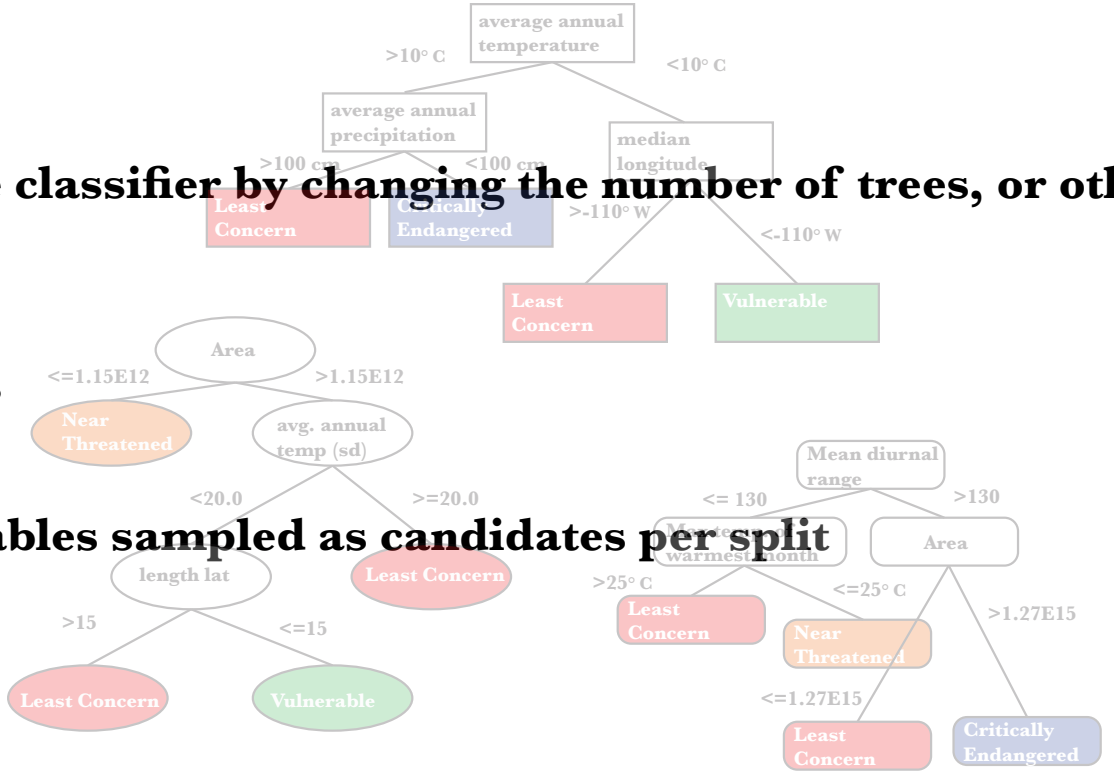
- **MSE for Regression Random Forest**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)$$

5. Improving the Classifier: Modeling Options

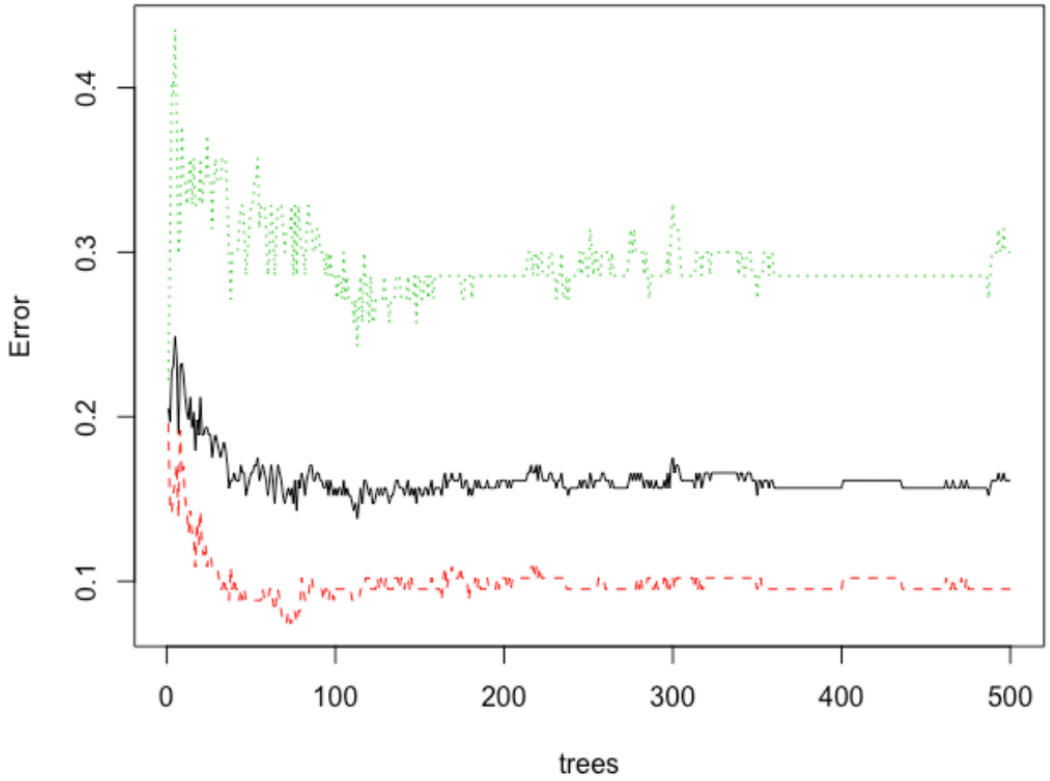
Can we improve the classifier by changing the number of trees, or other modeling options?

- **Number of trees**
- **Number of variables sampled as candidates per split**
- **Size of the tree**



Improving the Classifier: Modeling Options

Did we use enough decision trees in the classifier?

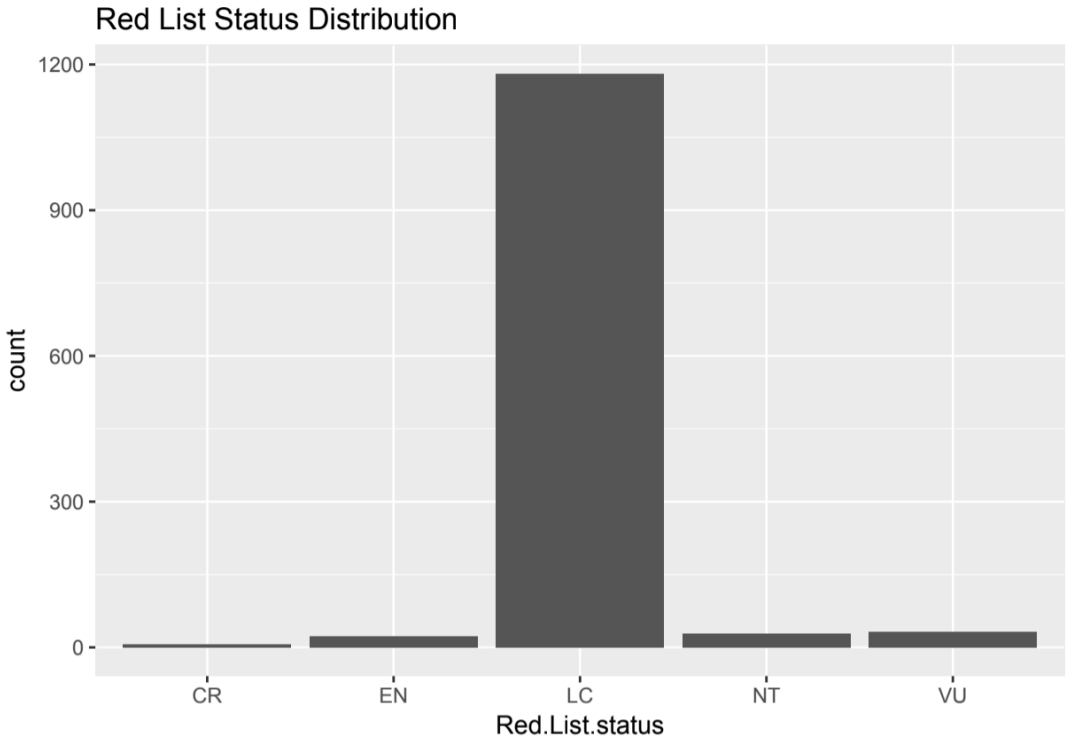


6 & 7. Improving the Classifier: Issues with Big Data

Is our dataset balanced? When classes are unbalanced, error rates may be inflated.

We can diminish this issue using a couple of approaches:

1. Down-sampling
2. Re-defining classes



8. Variable Importance:

Mean Decrease in Accuracy: the average increase in the number of times a case is OOB and misclassified when the variable is permuted compared to the number of OOB misclassifications when the variable is not permuted.

- Mean Decrease in Accuracy (MDA)
  1. train forest
  2. measure out-of-bag (OOB) accuracy (OOB.base)
  3. permute the values of variable *i* throughout training dataset
  4. Train forest with permuted variable *i* and measure OOB accuracy (OOB.permuted)
  5. *VarImp_i* = - (OOB_permuted – OOB.base)

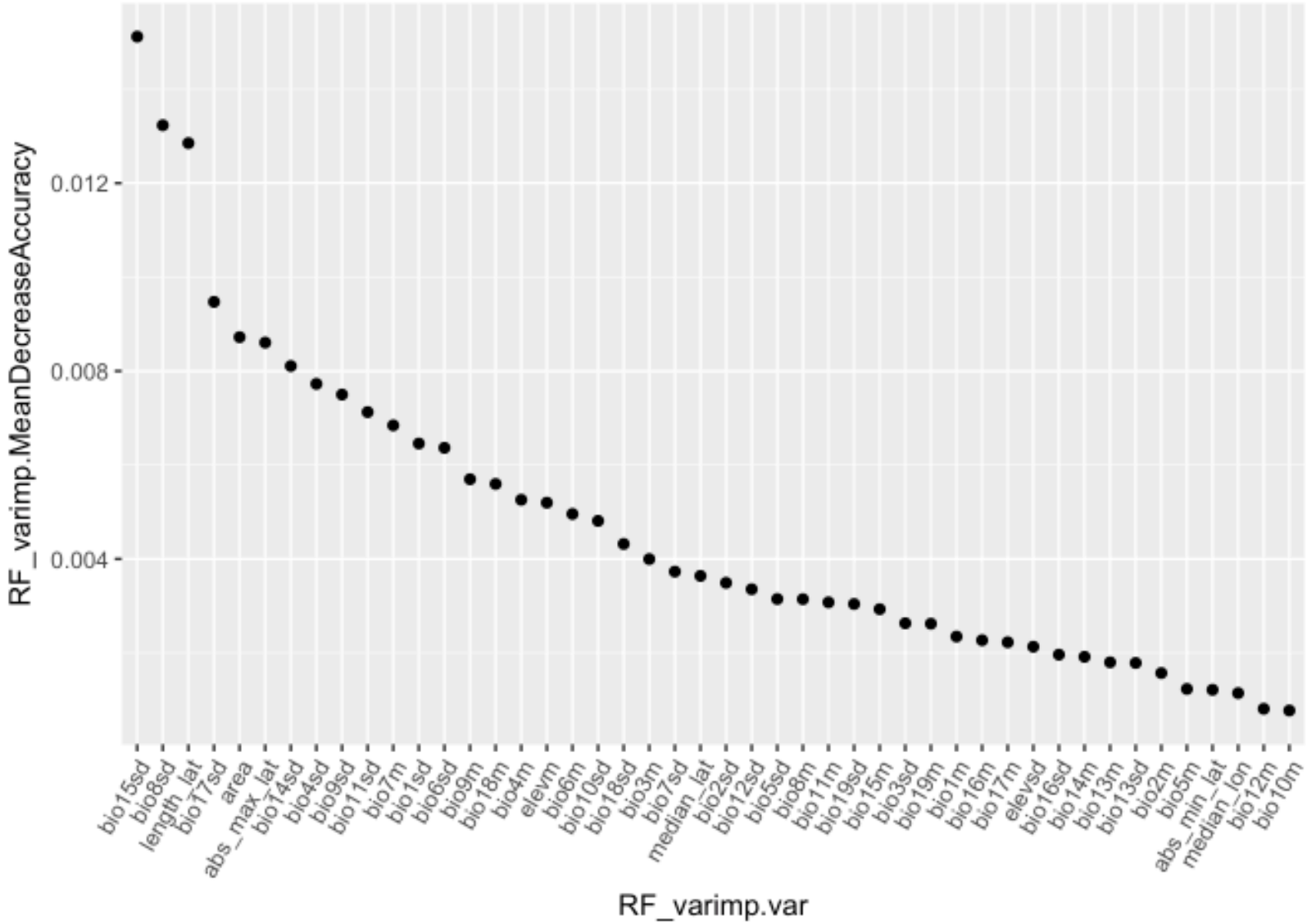
  - This is done across all trees, so for each variable you can calculate MDA as:

$$\frac{\text{Mean}(VarImp\_i)}{\text{Standard Deviation}(VarImp\_i)}$$

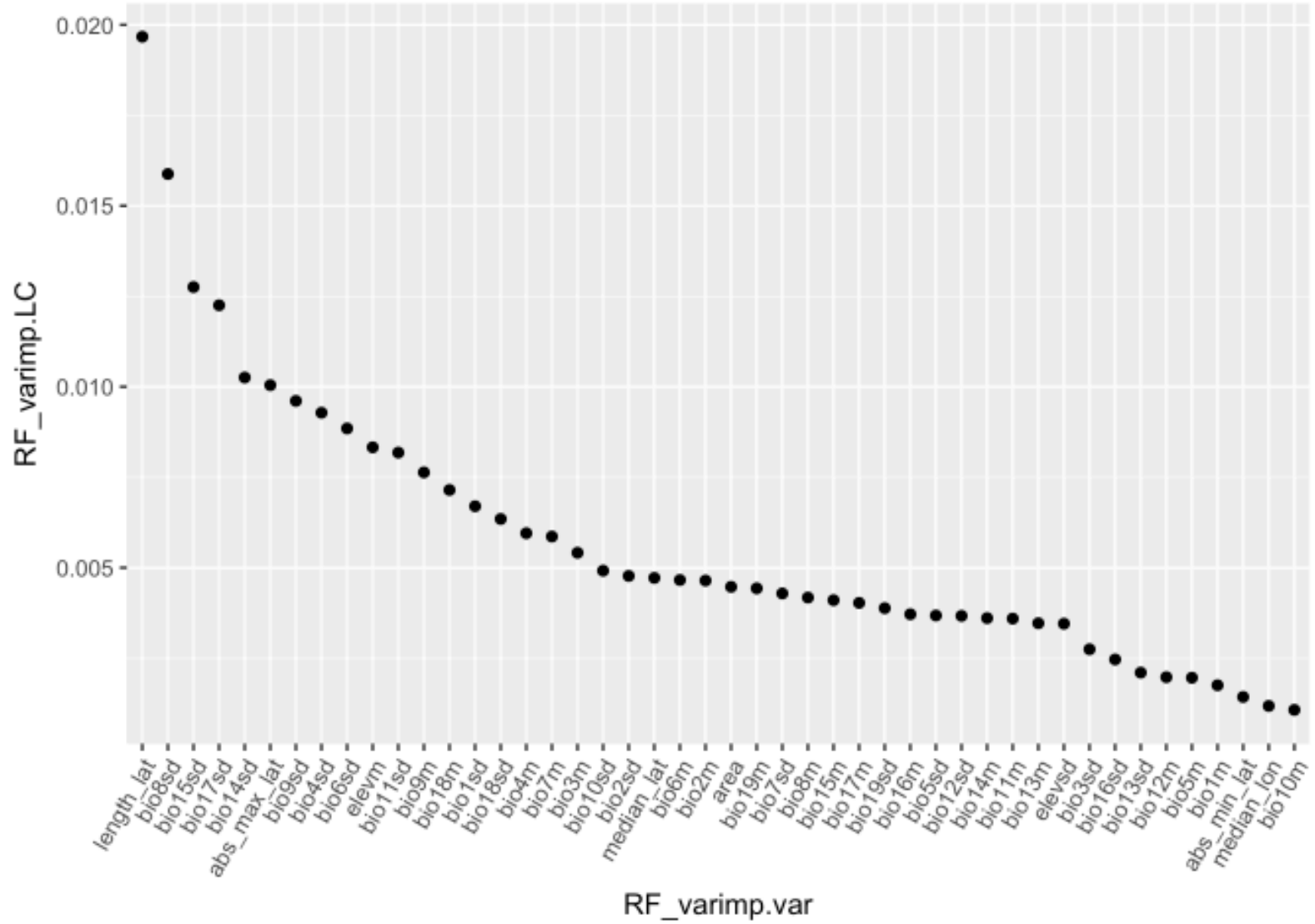
8. Variable Importance: GINI

- Gini impurity: measures variable importance based on how variables contribute to node purity
  - if a variable results in splits that generally split between, not within, classes, then that variable increases node purity
- Variables that increase node purity will have higher mean decreases in GINI.
- Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

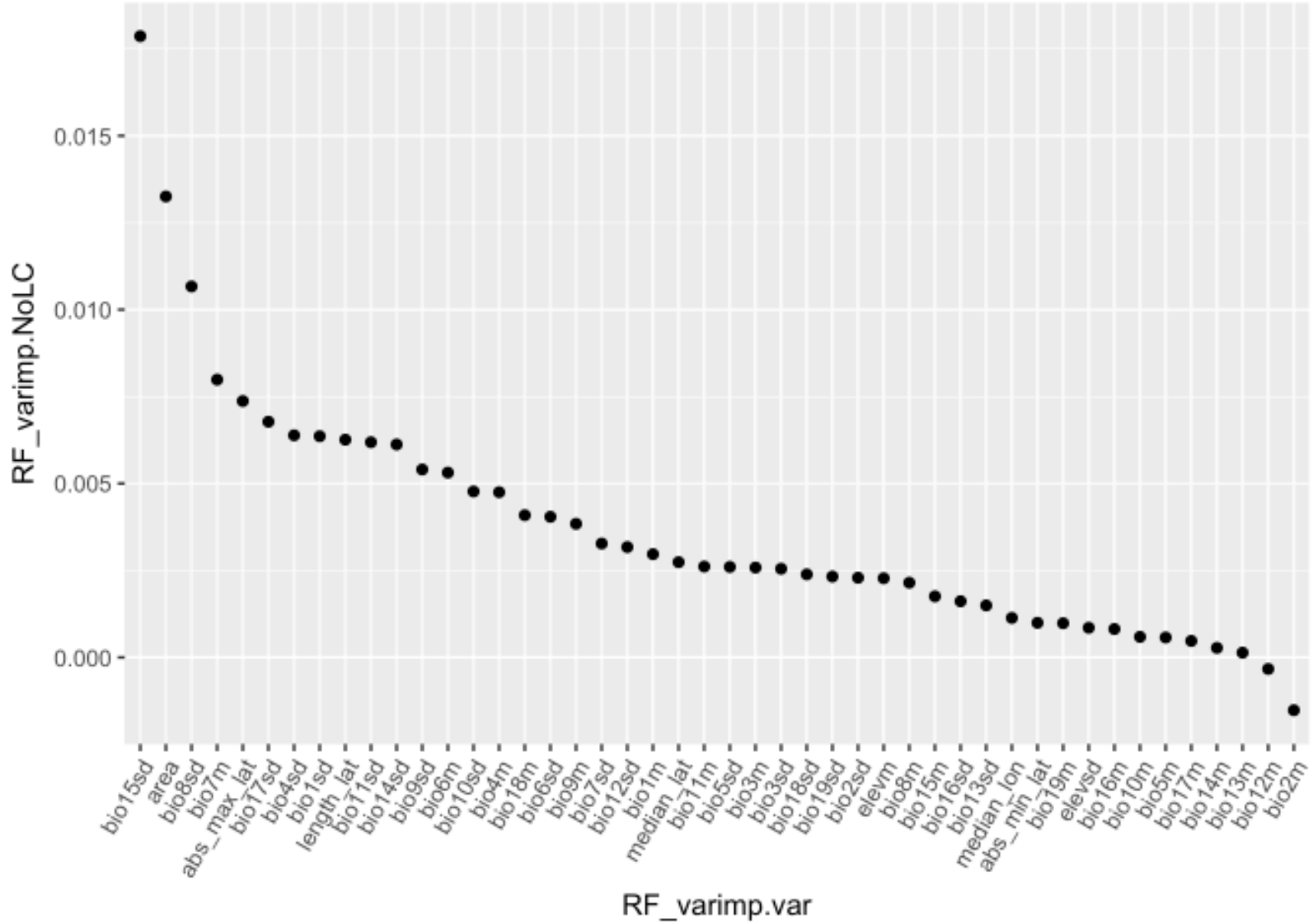
8. Variable Importance: *MDA entire forest*



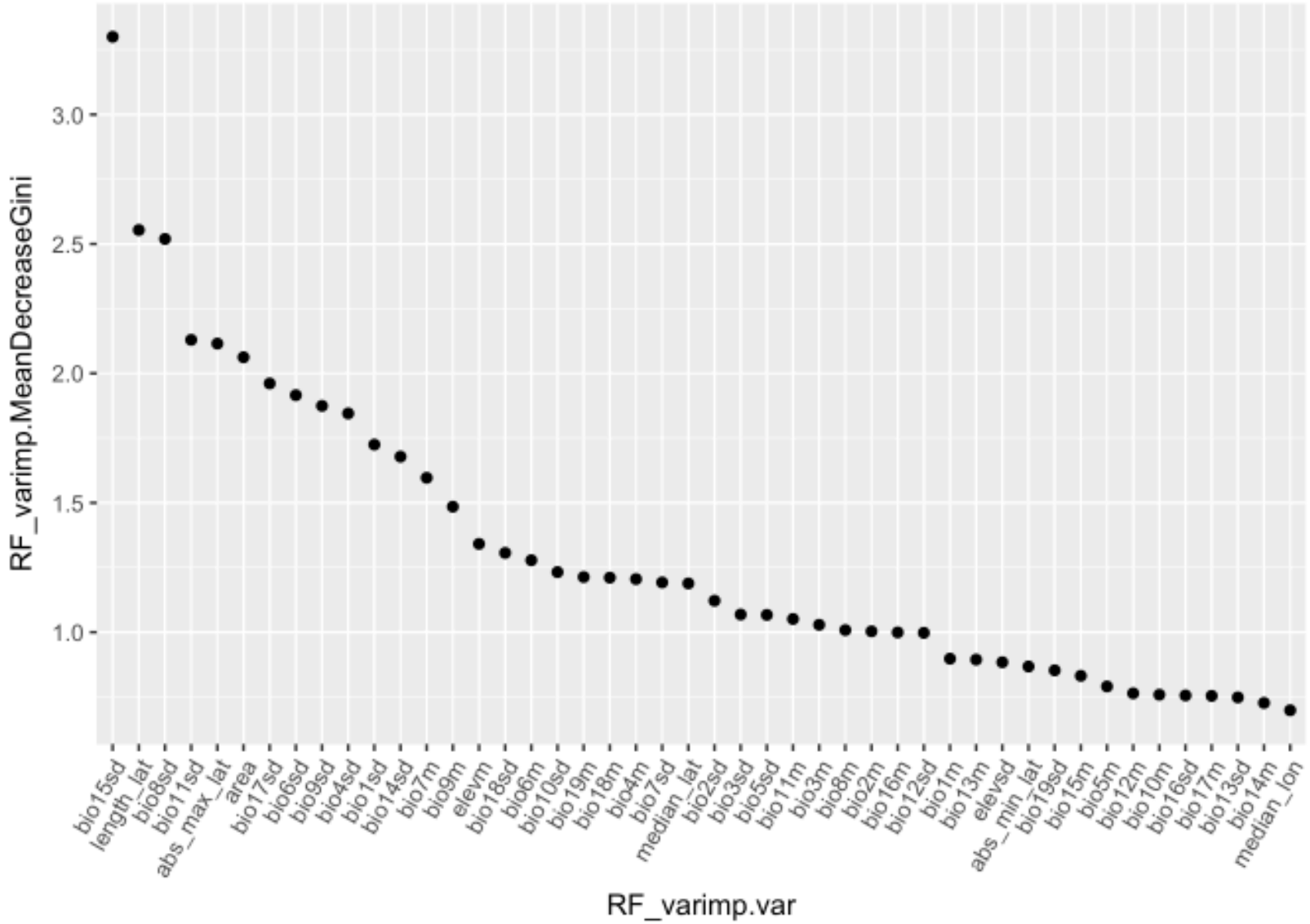
8. Variable Importance: *MDA LC class only*



8. Variable Importance: *MDA non LC class only*



8. Variable Importance: *Mean Decrease GINI entire forest*

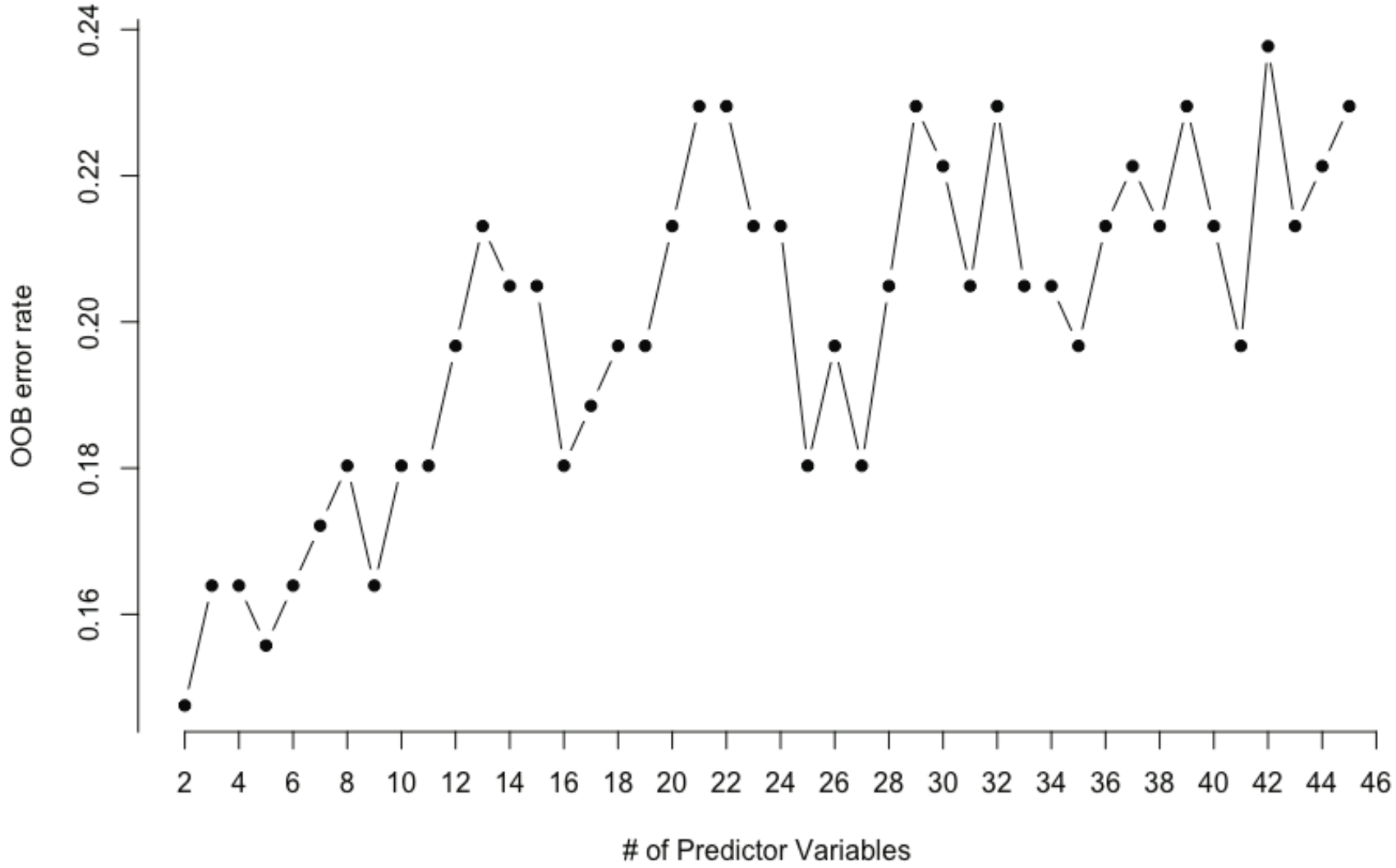


8. Variable Importance: *Can we remove the less informative predictor variables to improve our classifier?*

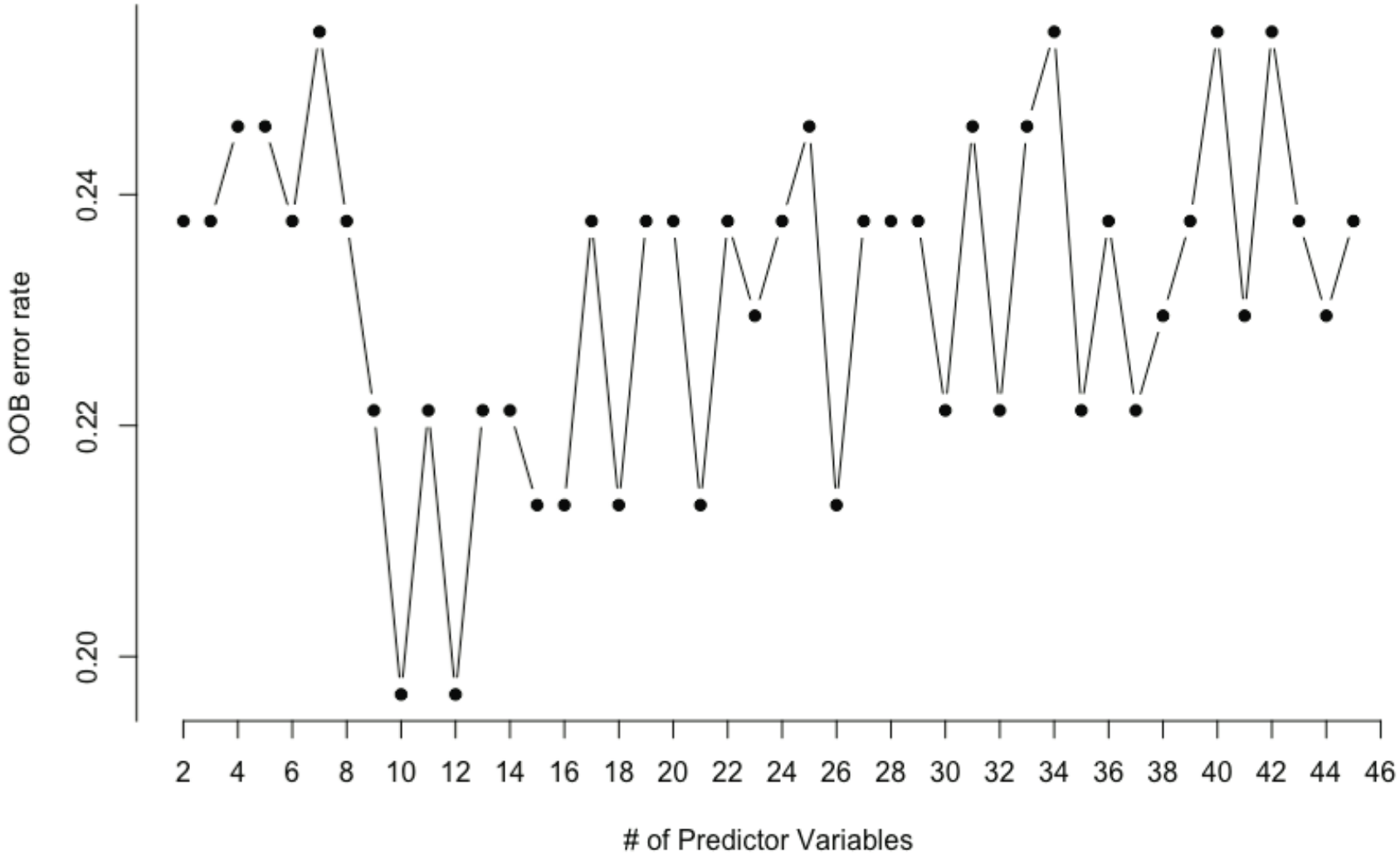
- Iteratively remove the least informative predictor variable (as determined by lowest MDA) and reconstruct a random forest classifier

- Does the OOB error rate improve with uninformative variables removed?

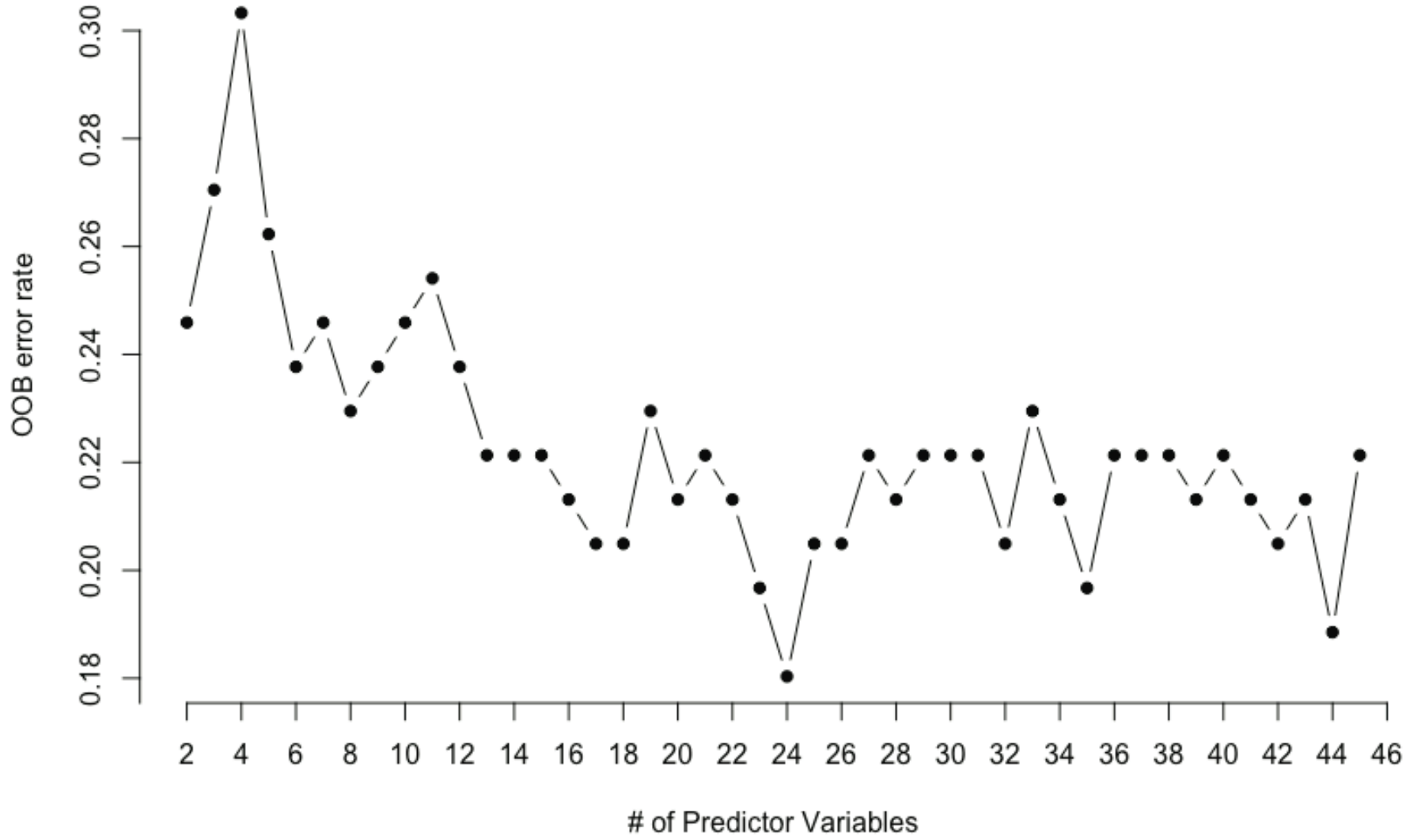
8. Variable Importance: *Can we remove the less informative predictor variables to improve our classifier?*



8. Variable Importance: *Can we remove the less informative predictor variables to improve our classifier?*



8. Variable Importance: *Can we remove the less informative predictor variables to improve our classifier?*



9. Predicting unknown responses:

- How good are the predictions?
- Do they generally support one class over the other?
- Is the support shared between both classes?

9. Predicting unknown responses:

Model with High Support	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Model with Low Support	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**	**0**