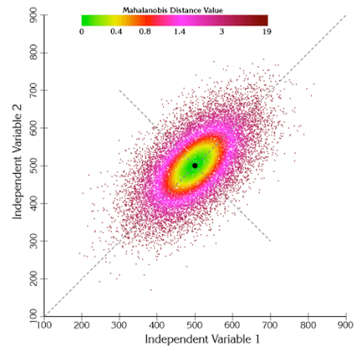
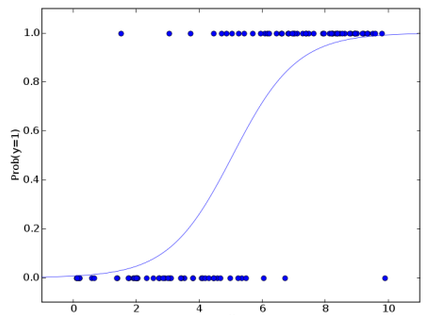
SDM techniques

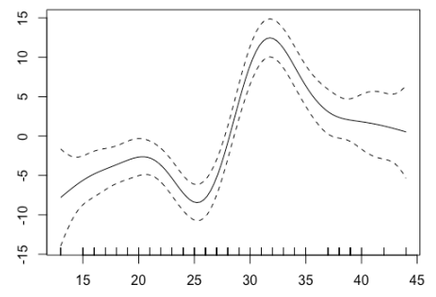
Statistical Models / Algorithms for modeling and mapping species distributions

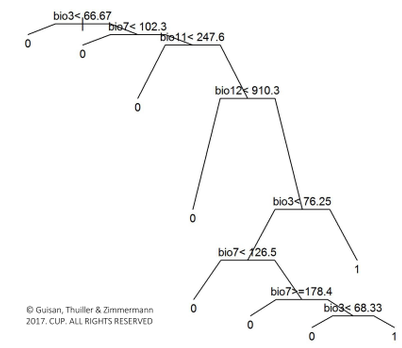
A few primary types

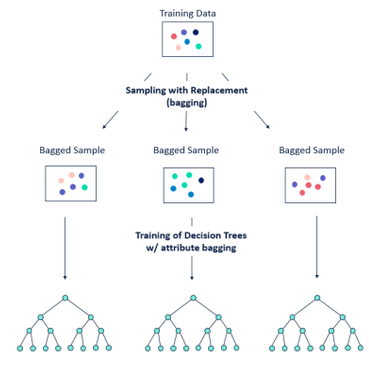
- Envelopes and Distance-based methods (simple category)
- Regression standard GLMs, GAMs
- Decision trees
- Machine Learning (ML) newest

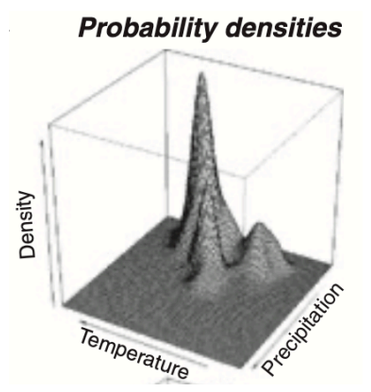


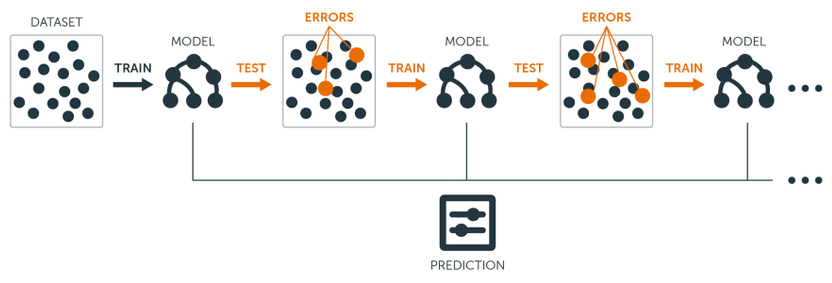








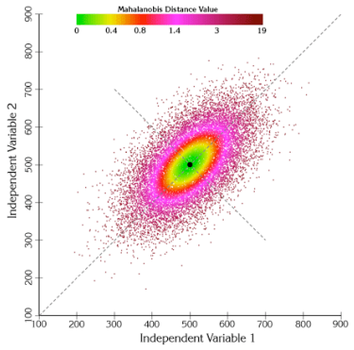


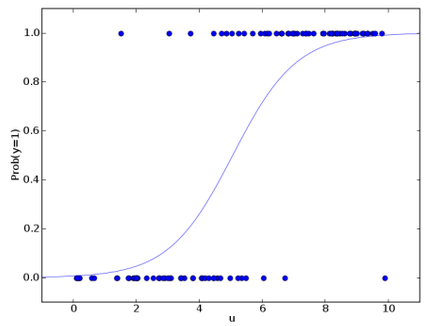


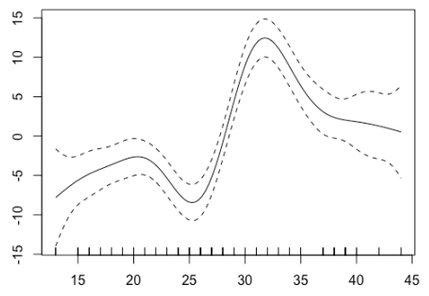
A few primary types

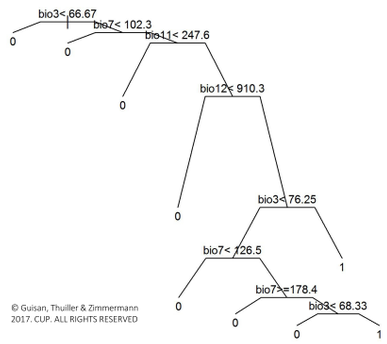
- Envelopes and Distance-based methods

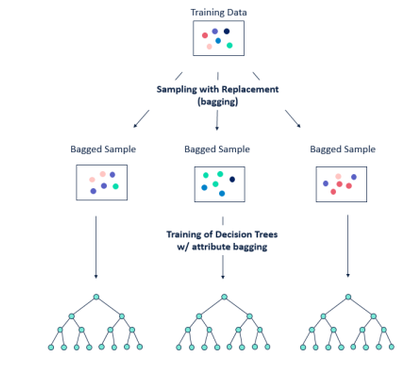
- BIOCLIM gets used the most & oldest
- DOMAIN
- **Mahalanobis distance**
- ENFA (Ecological Niche Factor Analysis)

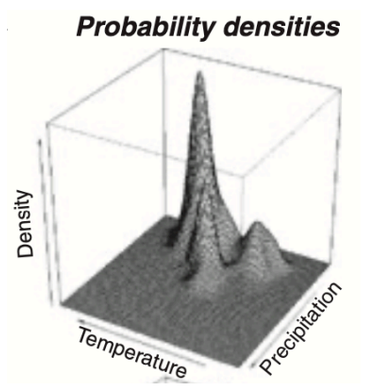


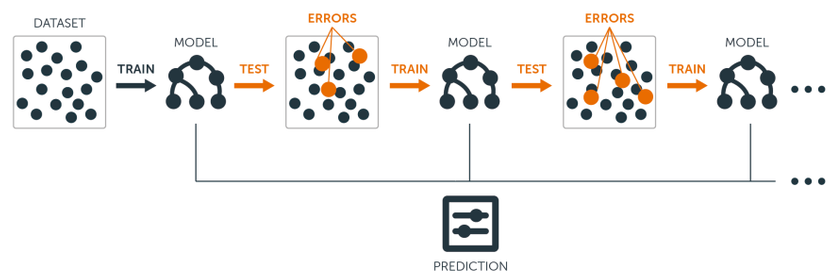






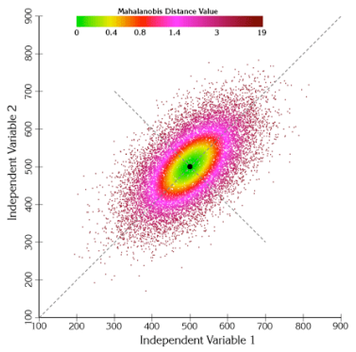


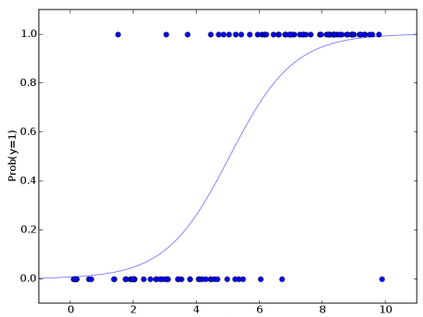


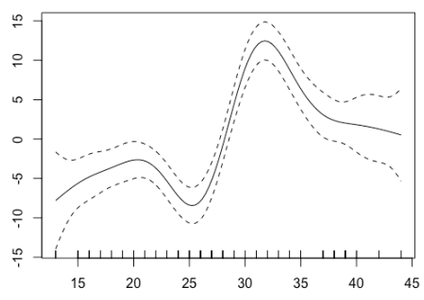


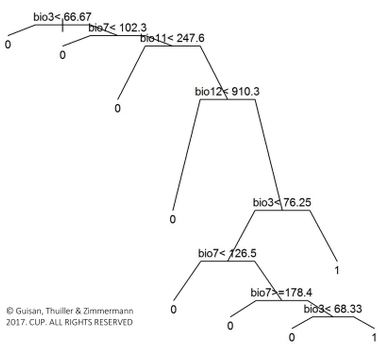
A few primary types

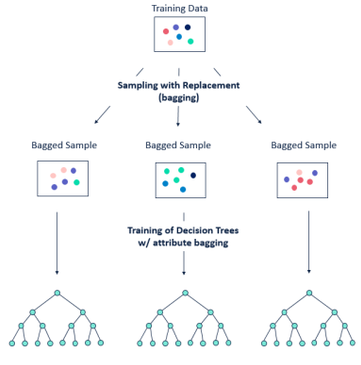
- Regression
  - **GLM (Generalized Linear Model)**
  - **GAM (Generalized Additive Model)**
  - MARS (Multivariate adaptive splines) used in spatial stats
  - Bayesian approaches used in Bayesian stats I think

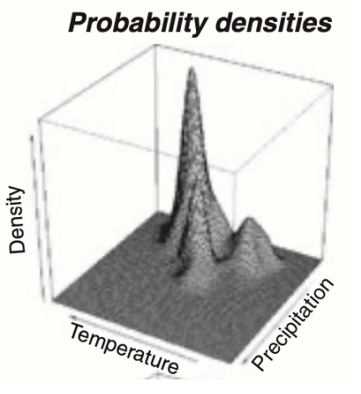


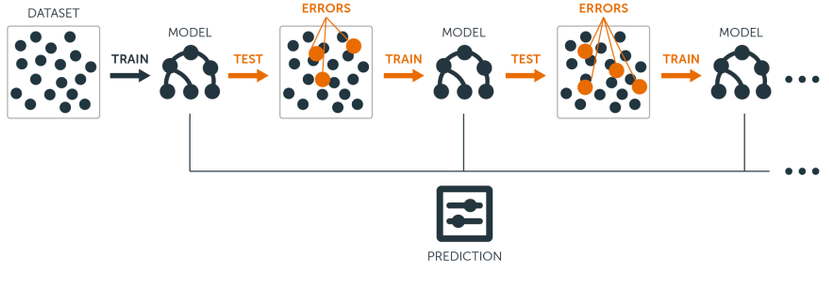






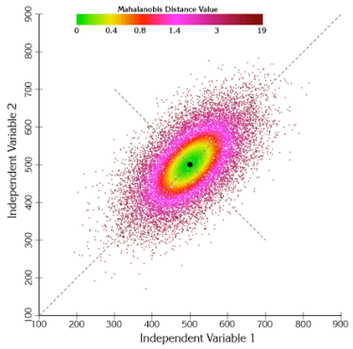


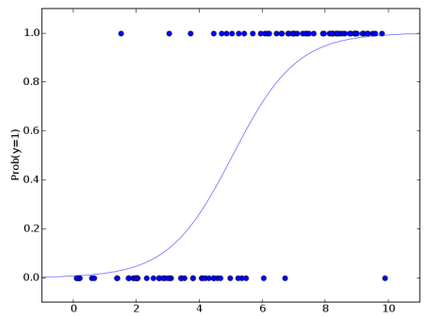


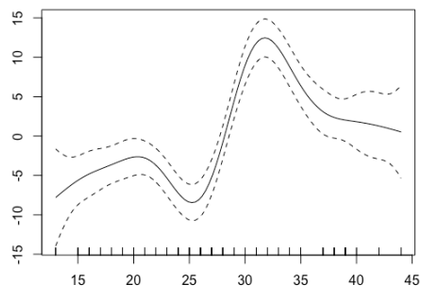


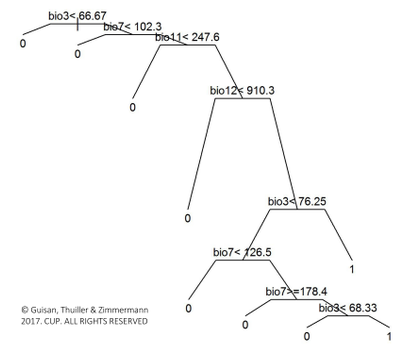
A few primary types

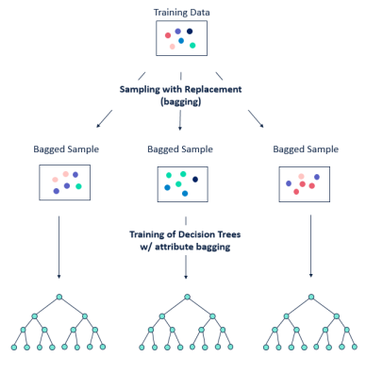
- Decision trees
  - CART (Classification / regression trees)

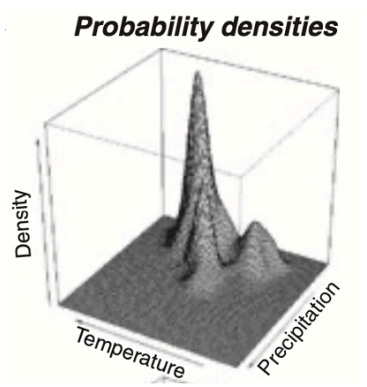


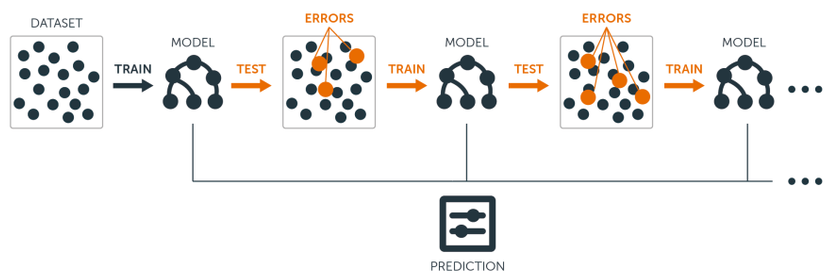








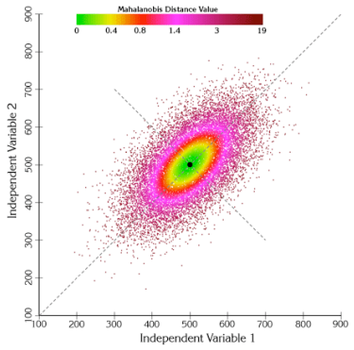


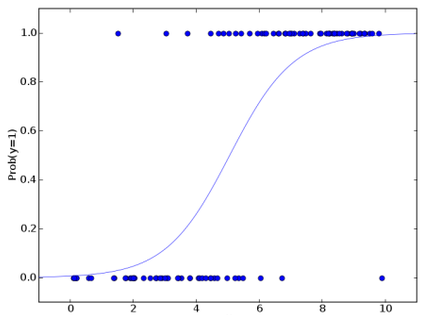


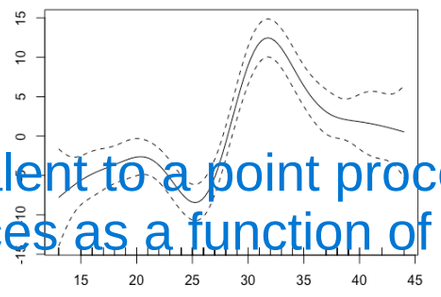
A few primary types

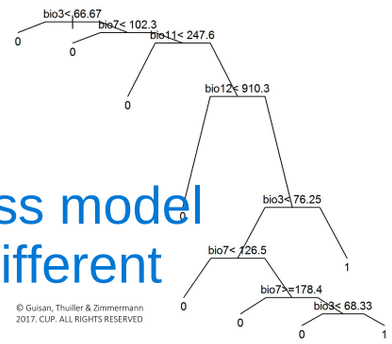
- Machine learning
  - ANN (artificial neural networks)
  - Genetic algorithms
  - **Maximum entropy**
  - SVM (Support Vector Machines) saw this in the SSDM package
  - **Random Forests**
    machine learning approaches
  - **Boosted Regression Trees (BRT/GBM)**

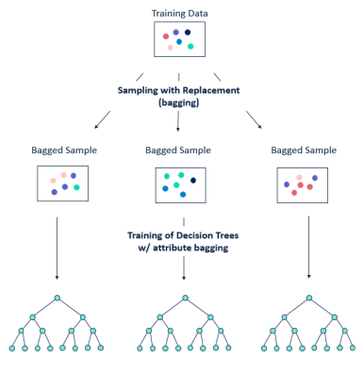
MAXENT basically equivalent to a point process model intensity species ocurrences as a function of different environmental covariates

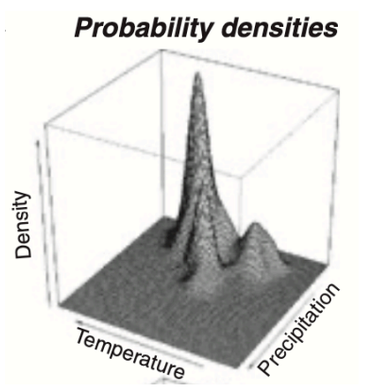


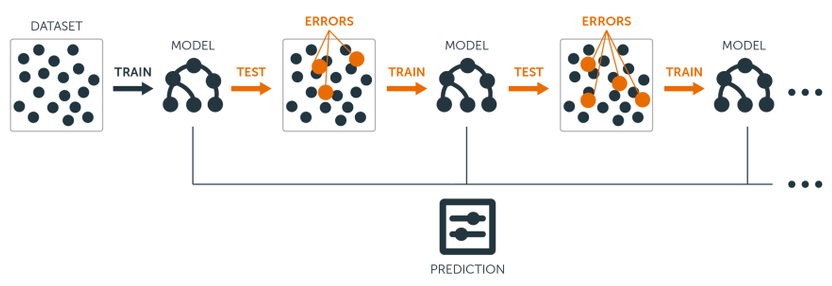










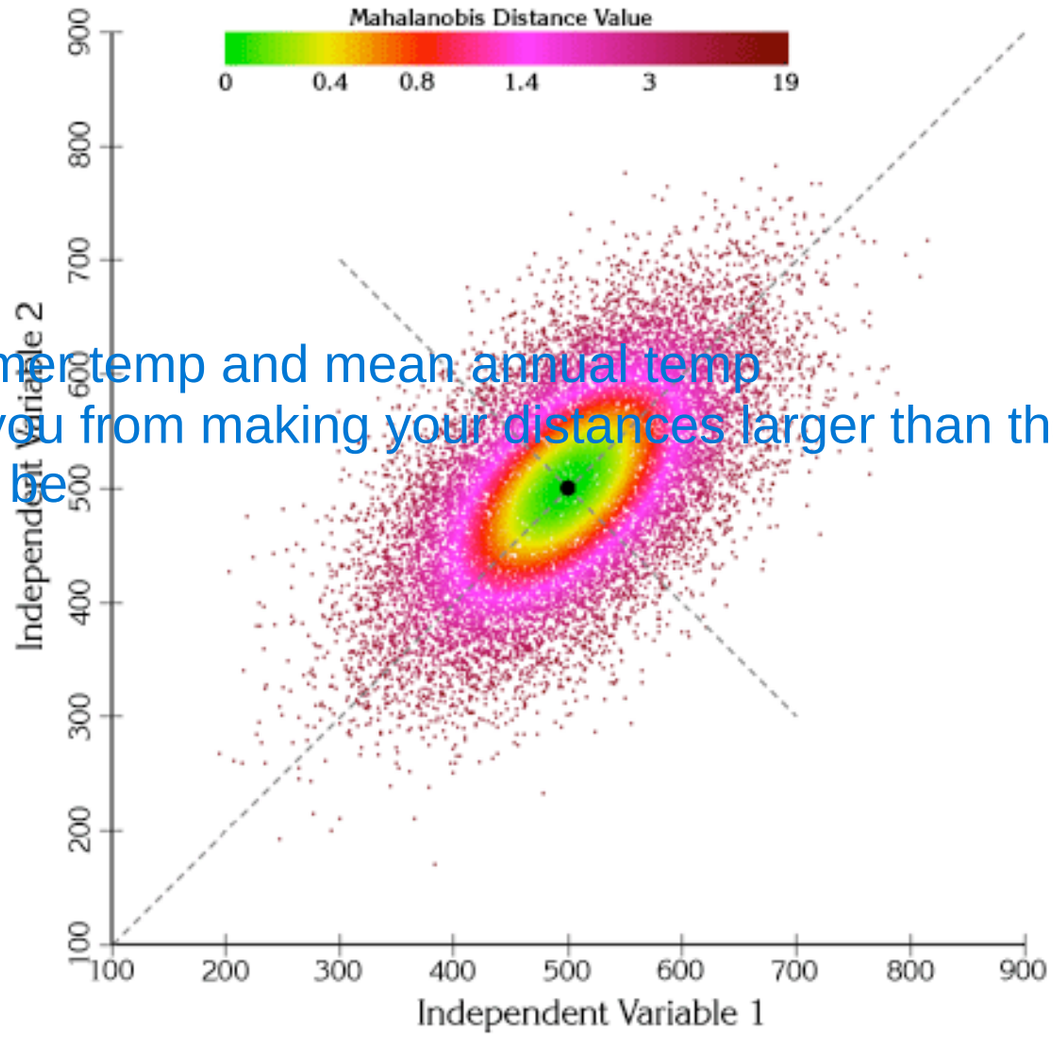


Mahalanobis distance

Presence ONLY method! Don't need absences or background data

given these other places that I want to predict to, how similar are they to this one?

- What is it?
  - A measure of multivariate similarity
  - Accounts for correlations between variables and their scale
  - Compares mean conditions of the species records to new observations (i.e., grid cells in region) & asks: how far are these new data from the cloud of observations?
  - Output is D^2, which can be converted to a p-value



ex: mean summer temp and mean annual temp
stops you from making your distances larger than they should be

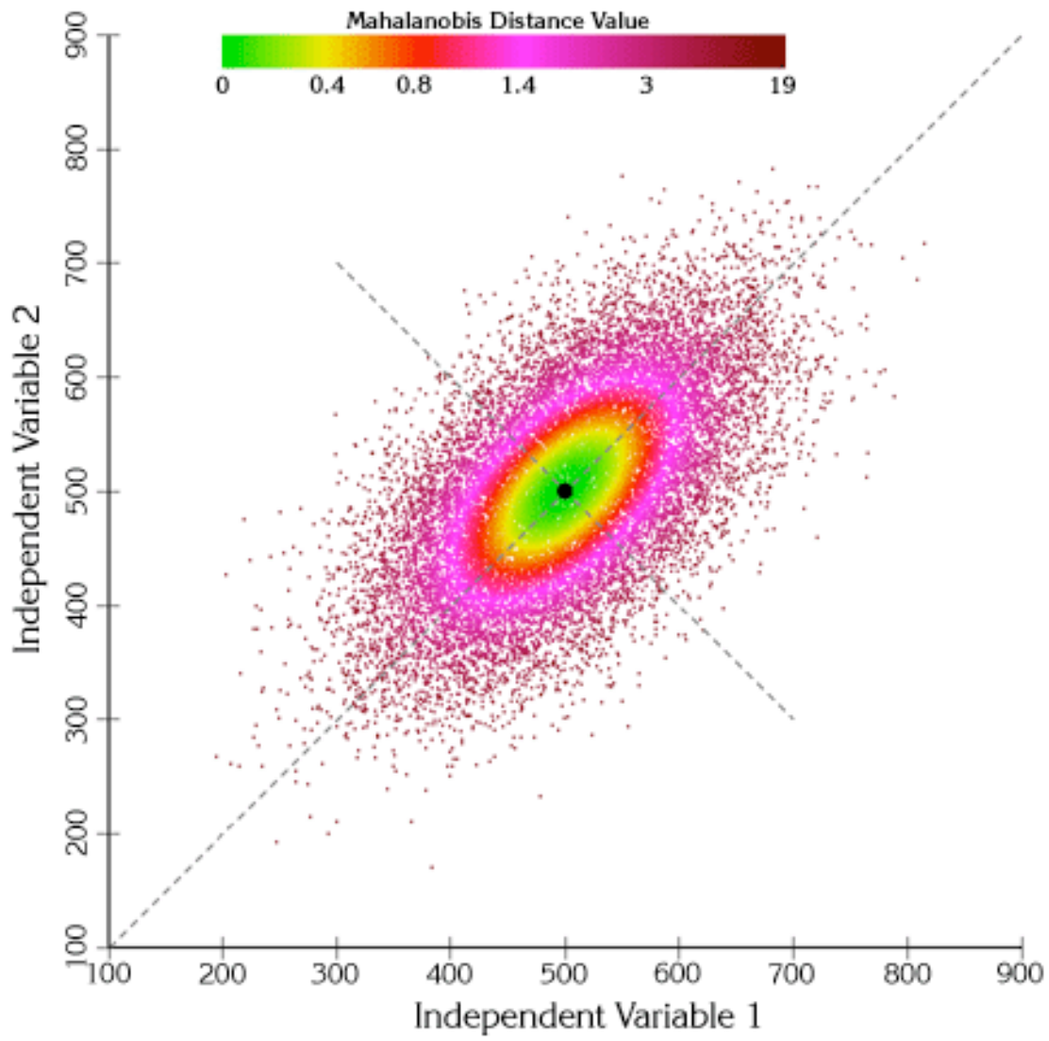
converts to a probability

within the R DISMO package

doesn't weight the variables' importance
doesn't allow for categorical variables

Mahalanobis distance

- Important points
  - A simple, accessible, presence-only method that tends to perform well
  - Collinearity is not an issue
  - Linear method in which variables have equal weight
  - Continuous predictors only
  - ‘mahal’ function in dismo

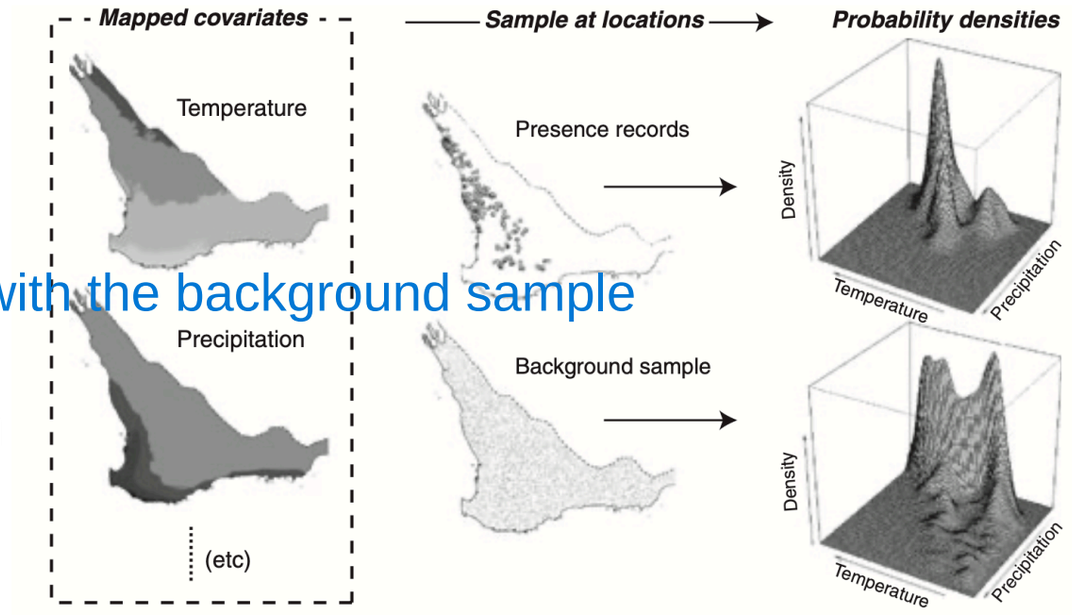


Maximum Entropy (MaxEnt)

Merow paper is good for extra info
https://doi.org/10.1111/j.1600-0587.2013.07872.x or https://doi.org/10.1111/geb.12453

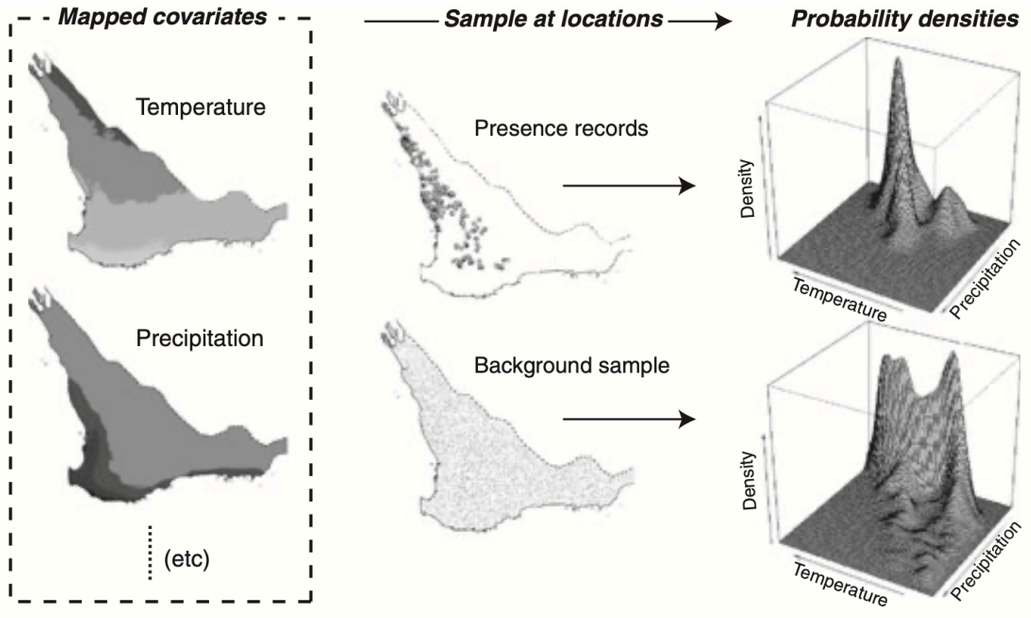
uses a random sample of the broader region to fit the model

- Presence-background density estimation method
- Estimates the multivariate distribution of suitable habitat conditions based on species occurrences and how those contrast with the background sample
- The best approximation of this distribution is the one with “maximum entropy” = the distribution that is most “spread out” subject to known **constraints**
- These **constraints** are defined by the expected value of the distribution estimated from the set of species occurrences



Maximum Entropy (MaxEnt)

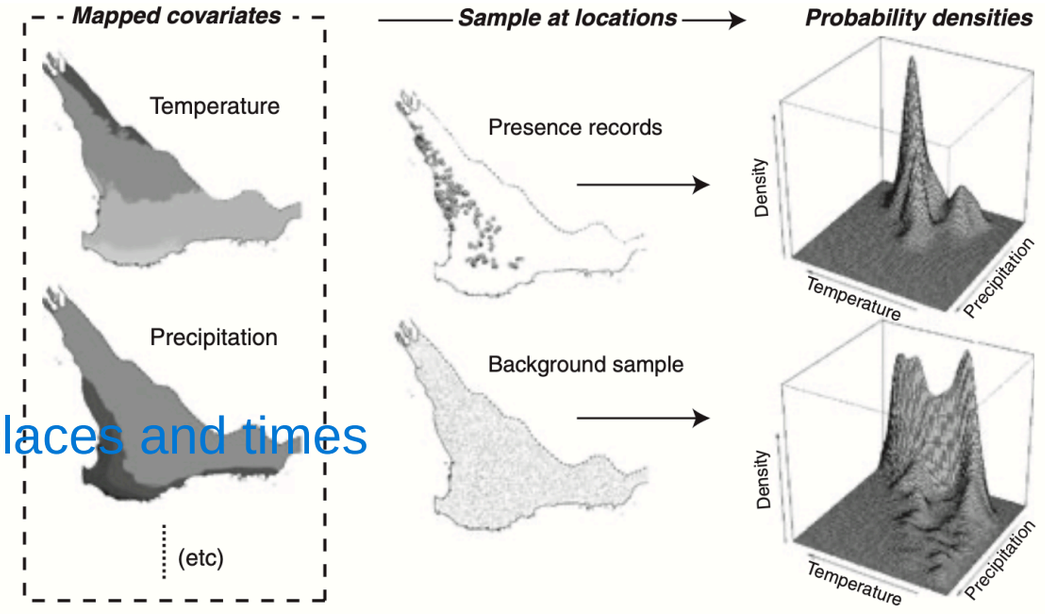
- What is maximum entropy in plain terms?
- It asks: how does the probability density of species records in environmental (covariate) space compare the probability density of the environment across the entire study region? compares in a multivariate manner



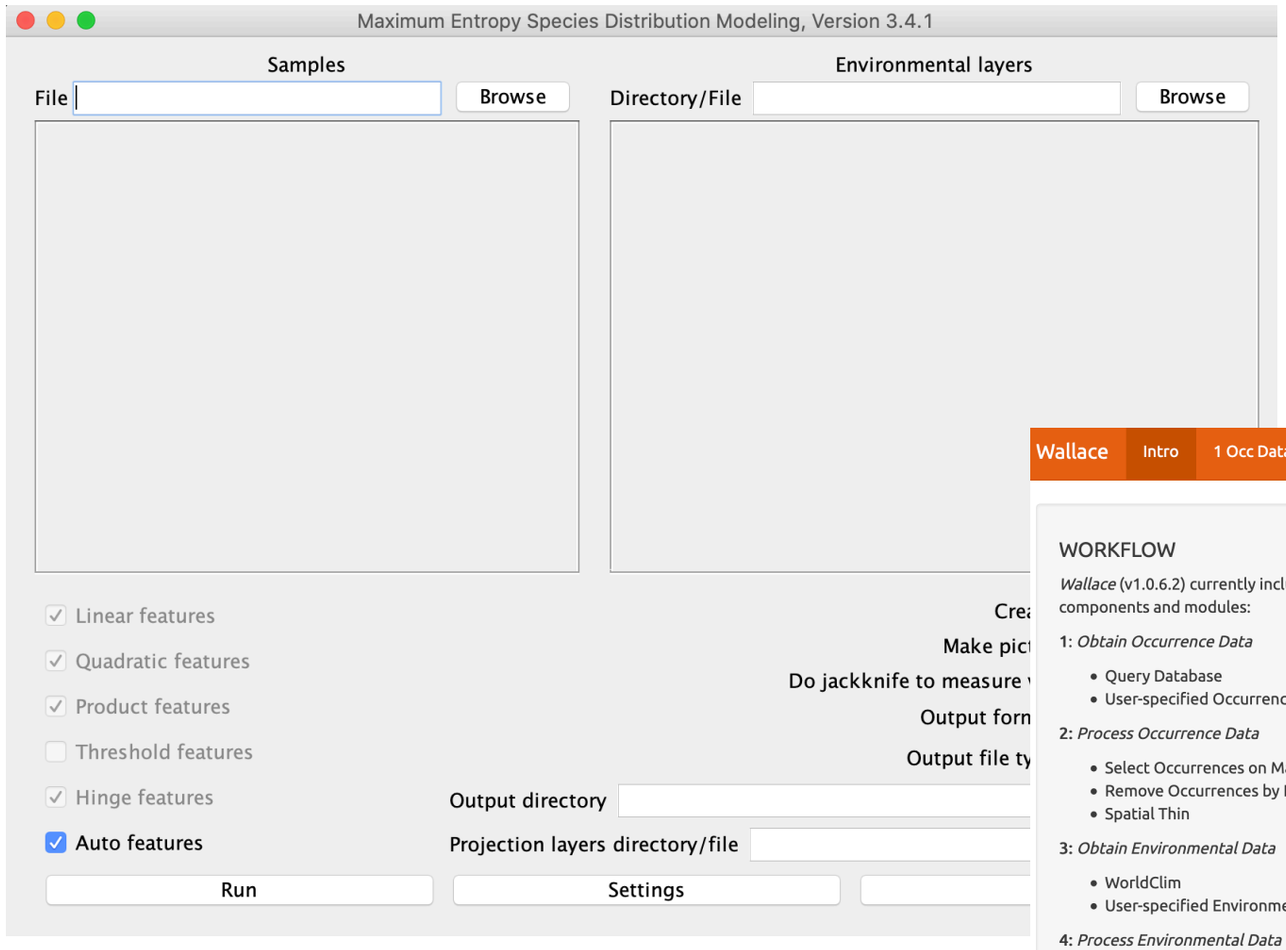
Maximum Entropy (MaxEnt)

- Advantages
  - A presence-only method that performs well
  - Robust to small sample sizes
  - Tries to avoid overfitting (but user can control) this is good for extrapolating to new places and times
- Disadvantages
  - User should make a number of decisions to which maxent can be highly sensitive
  - Software is user-friendly, but often there is not a good basis for selecting certain parameters.

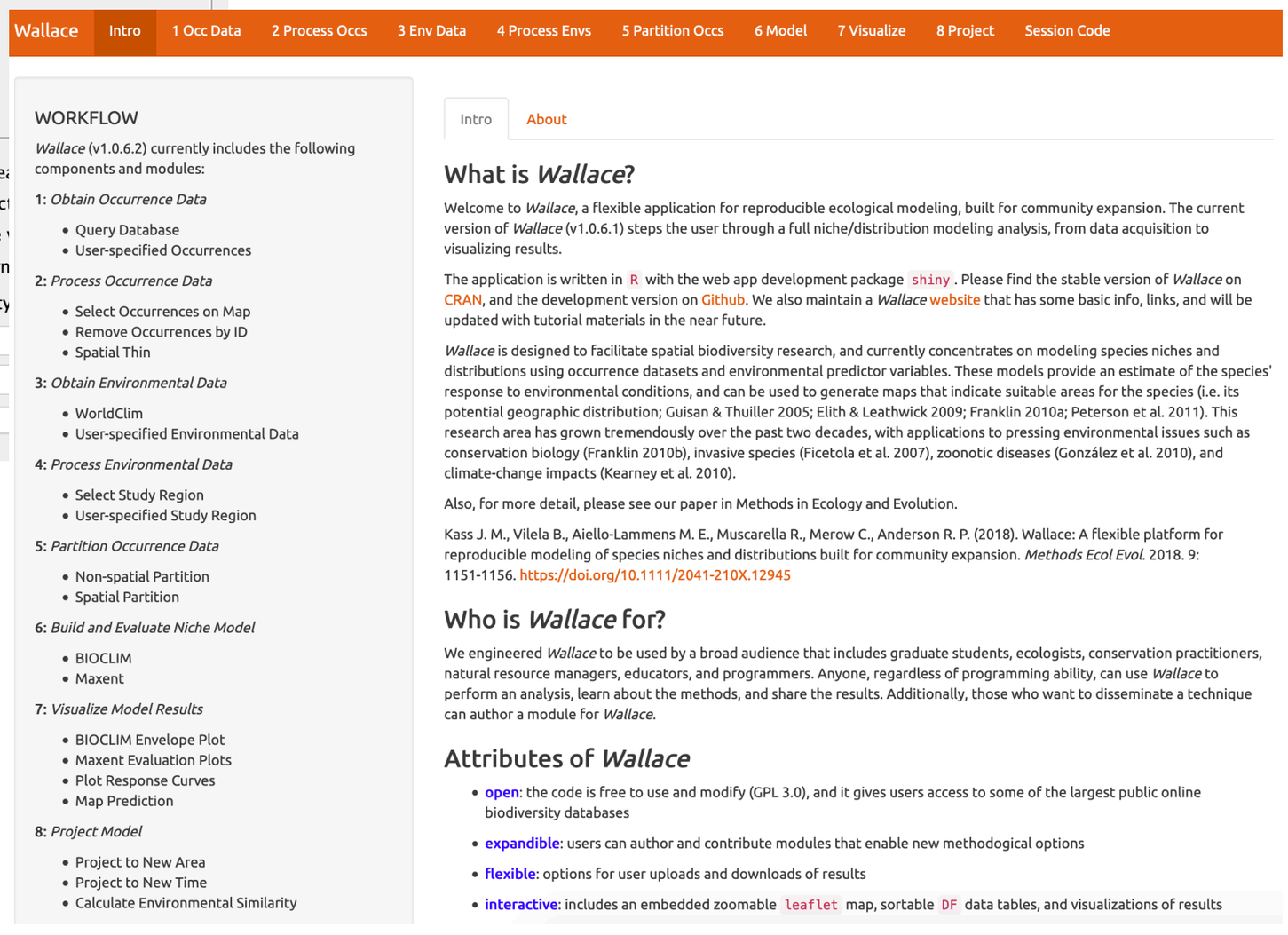
you should not just use the default settings-- the model is highly sensitive to user decisions of parameters



Maximum Entropy (MaxEnt)

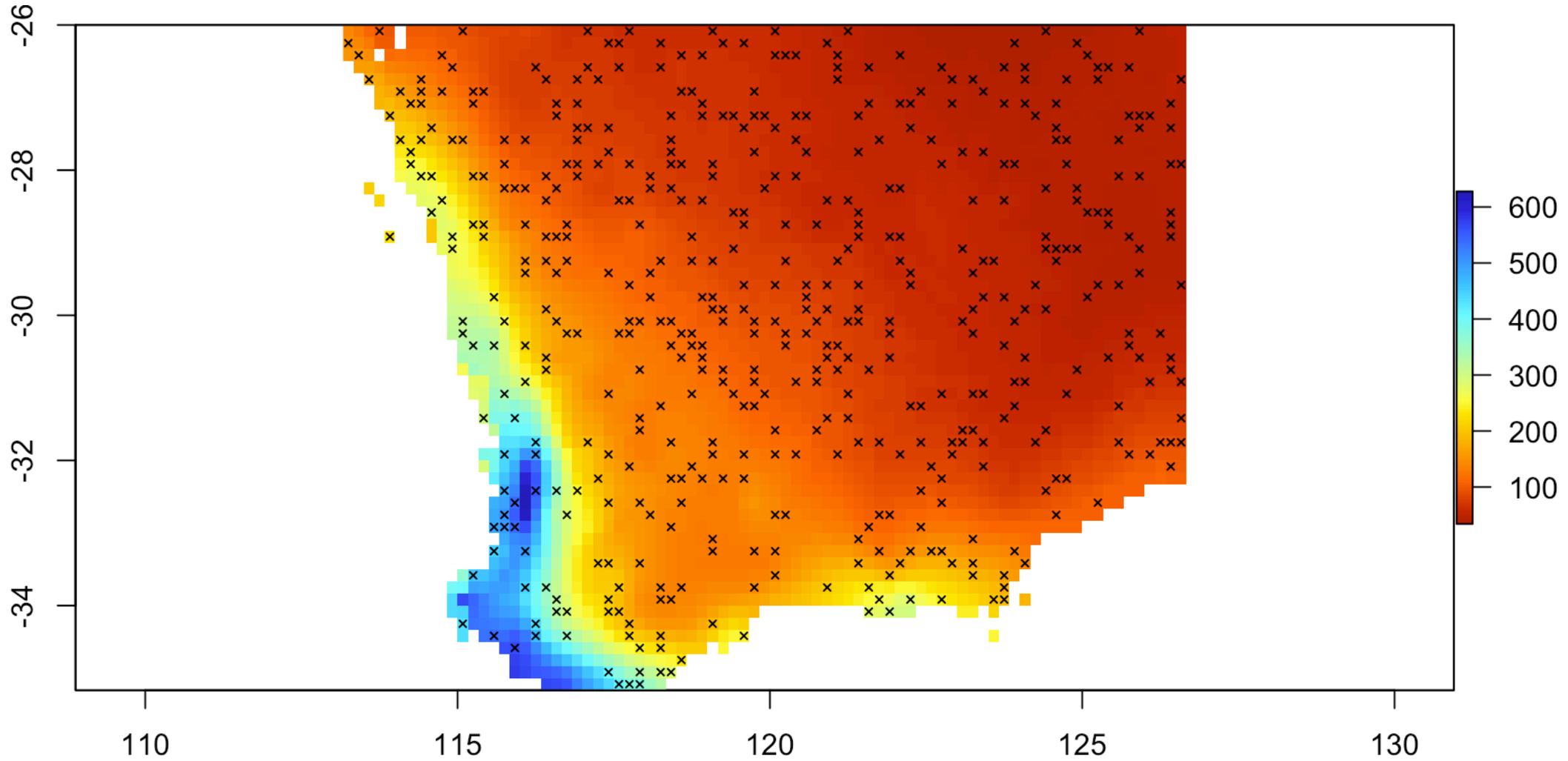


the stand alone program



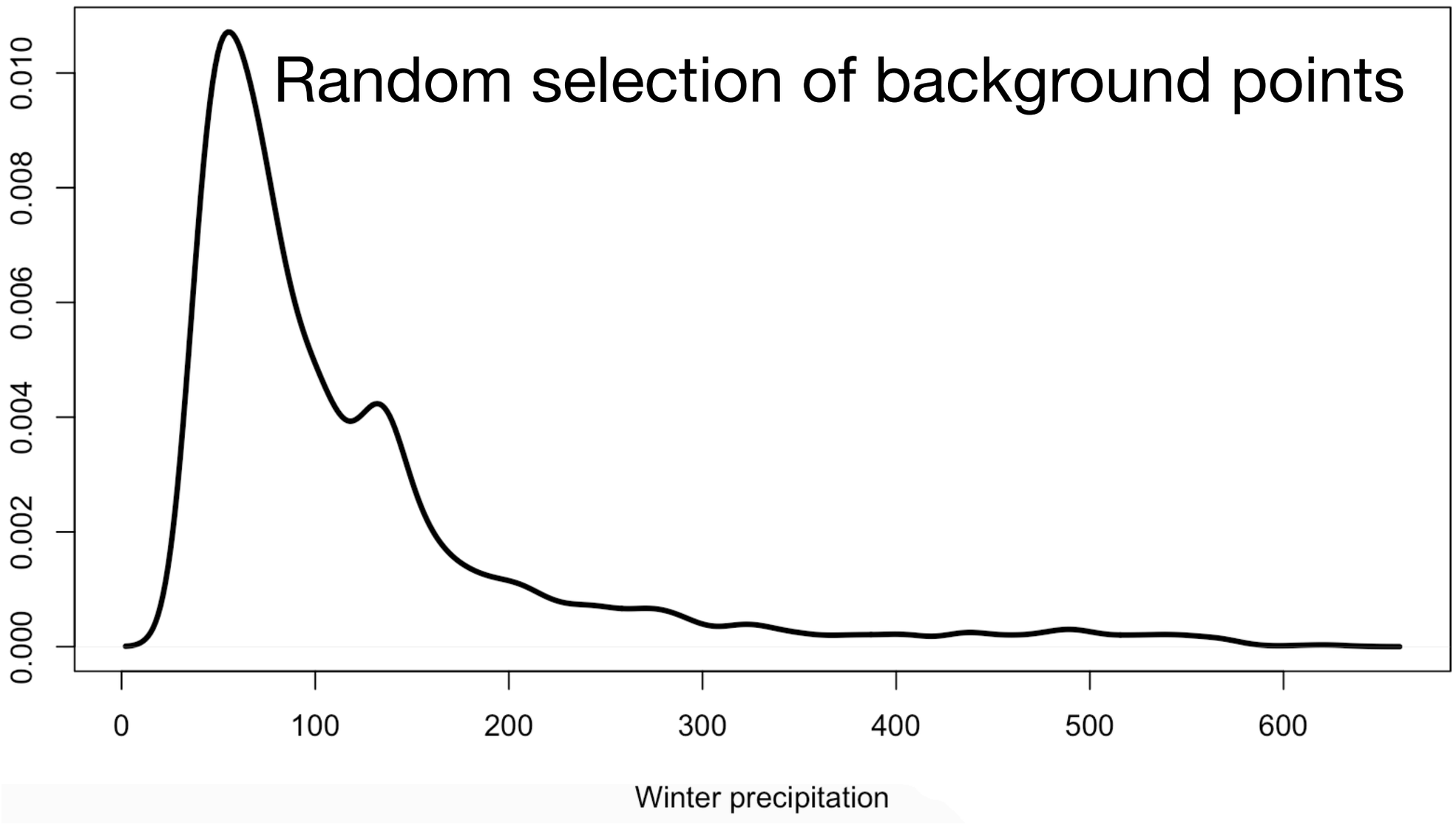
Maximum Entropy (MaxEnt)

default is sample of 10,000

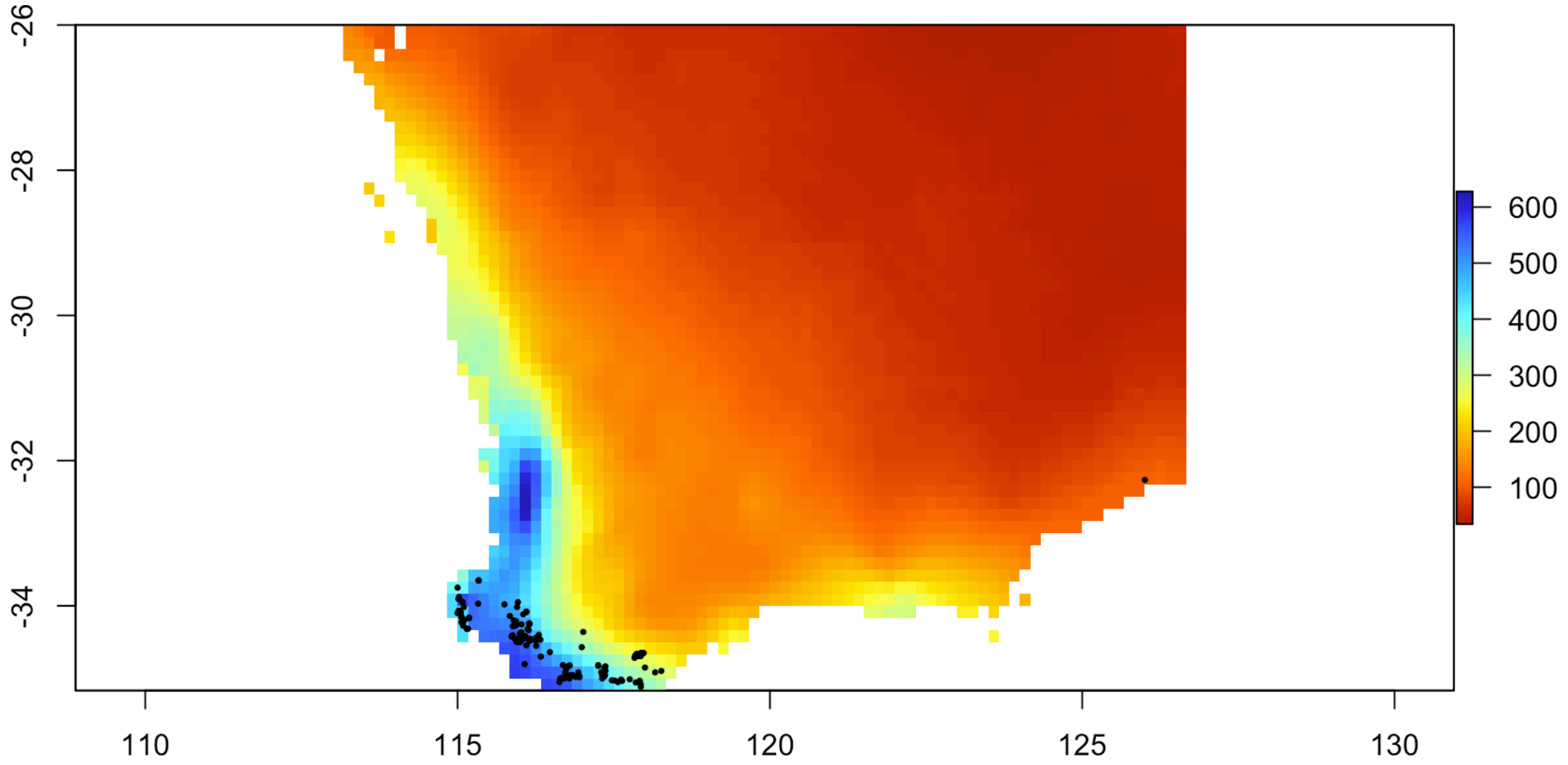


Random selection of background points

Maximum Entropy (MaxEnt)

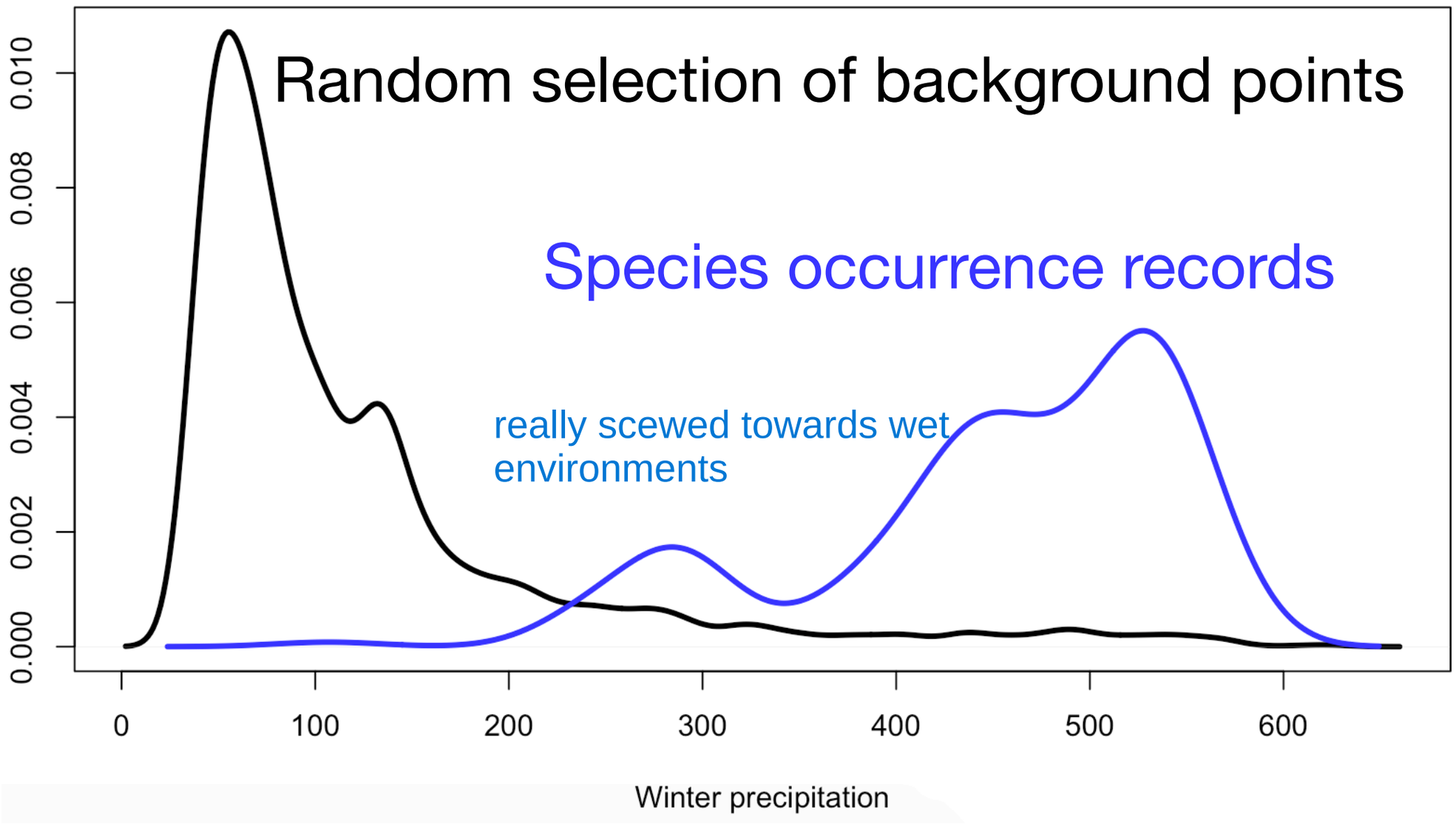


Maximum Entropy (MaxEnt)



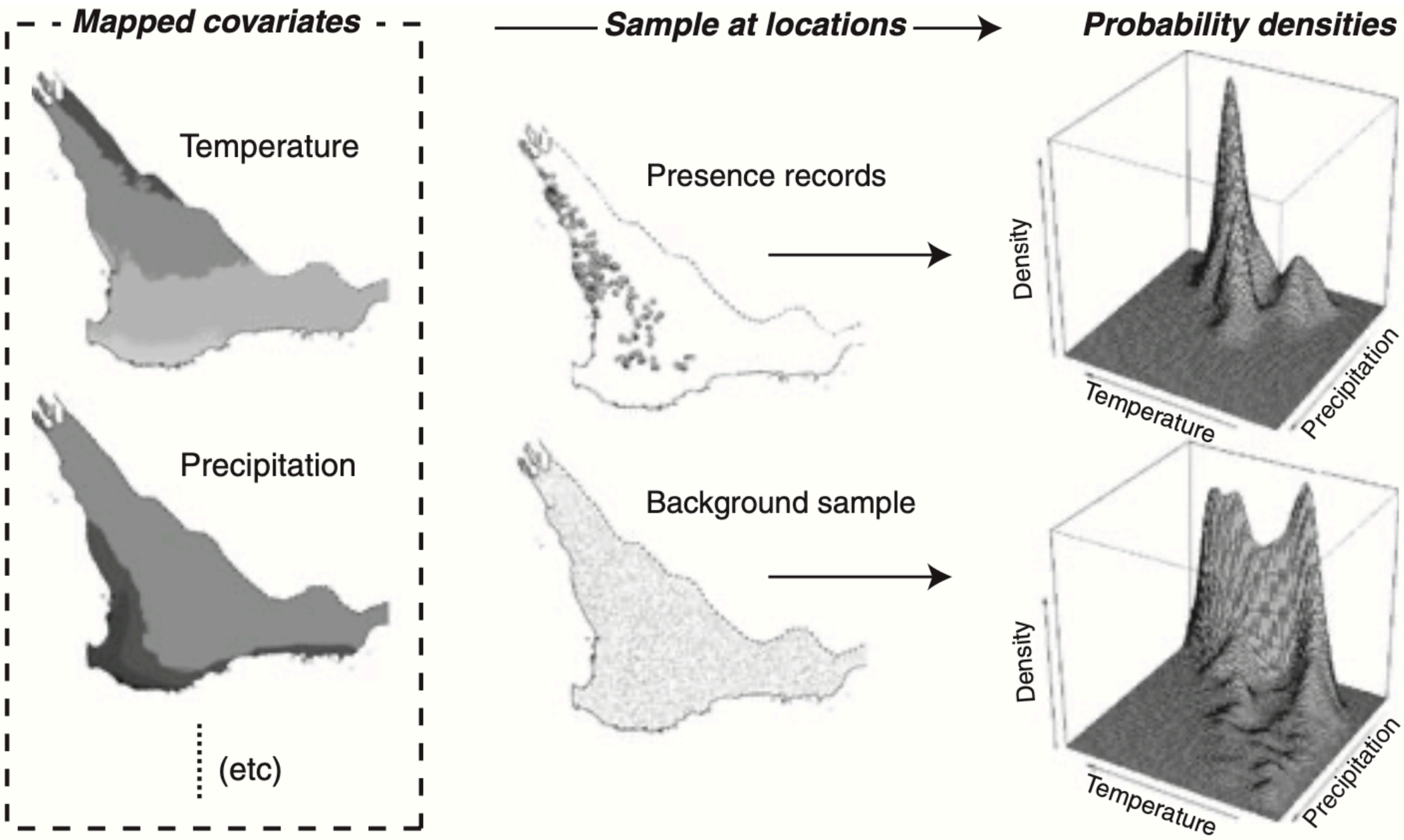
Species occurrence records

Maximum Entropy (MaxEnt)



Maximum Entropy (MaxEnt)

does this with all of the variables, not just the one variable above



Diversity and Distributions, (Diversity Distrib.) (2011) **17**, 43–57

BIODIVERSITY RESEARCH

A statistical explanation of MaxEnt for ecologists

Jane Elith[1]*, Steven J. Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1] and Colin J. Yates[5]

Maximum Entropy (MaxEnt)

Forum

EDITOR'S CHOICE

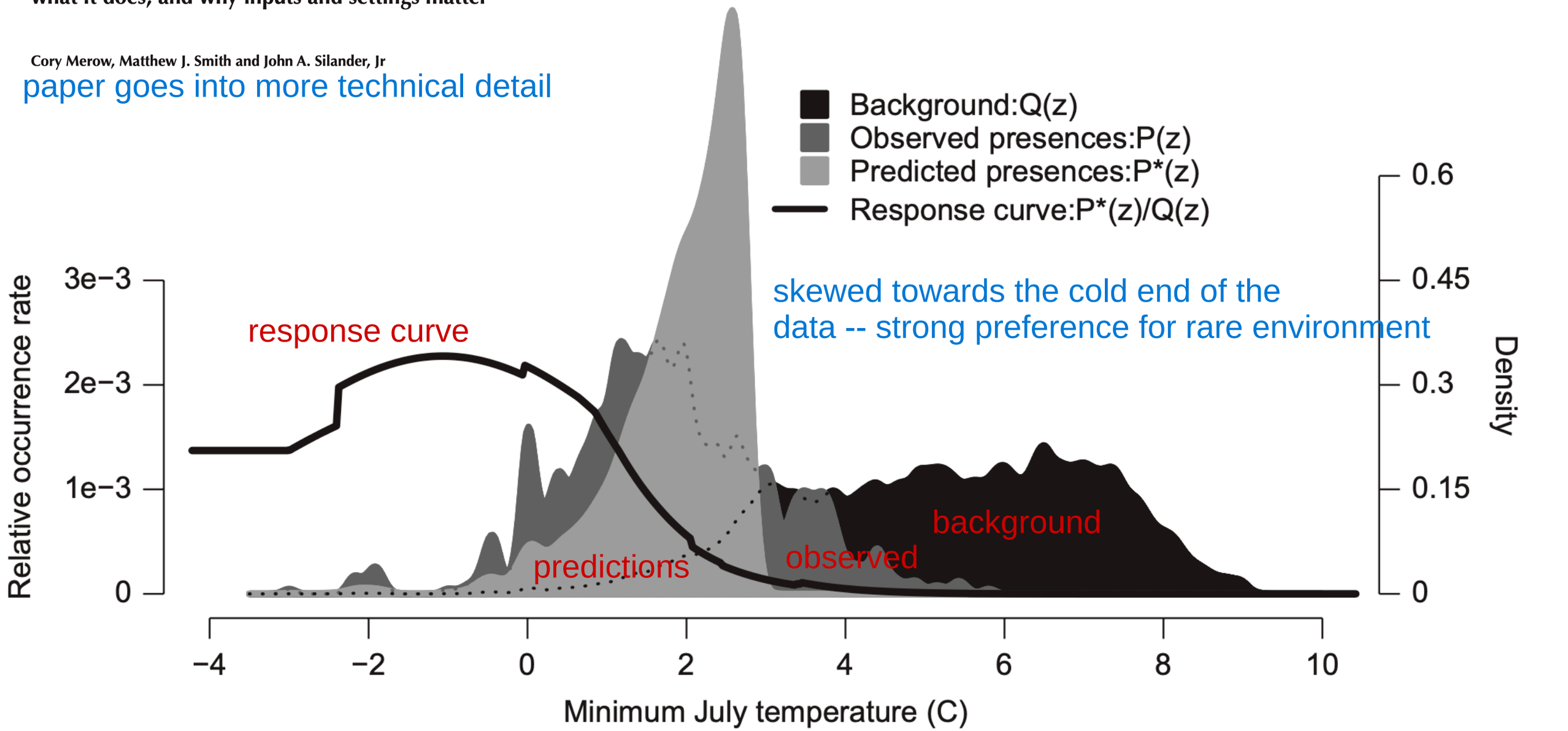
Ecography 36: 1058–1069, 2013
doi: 10.1111/j.1600-0587.2013.07872.x

Subject Editor: Niklaus E. Zimmermann. Accepted 27 March 2013

A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter

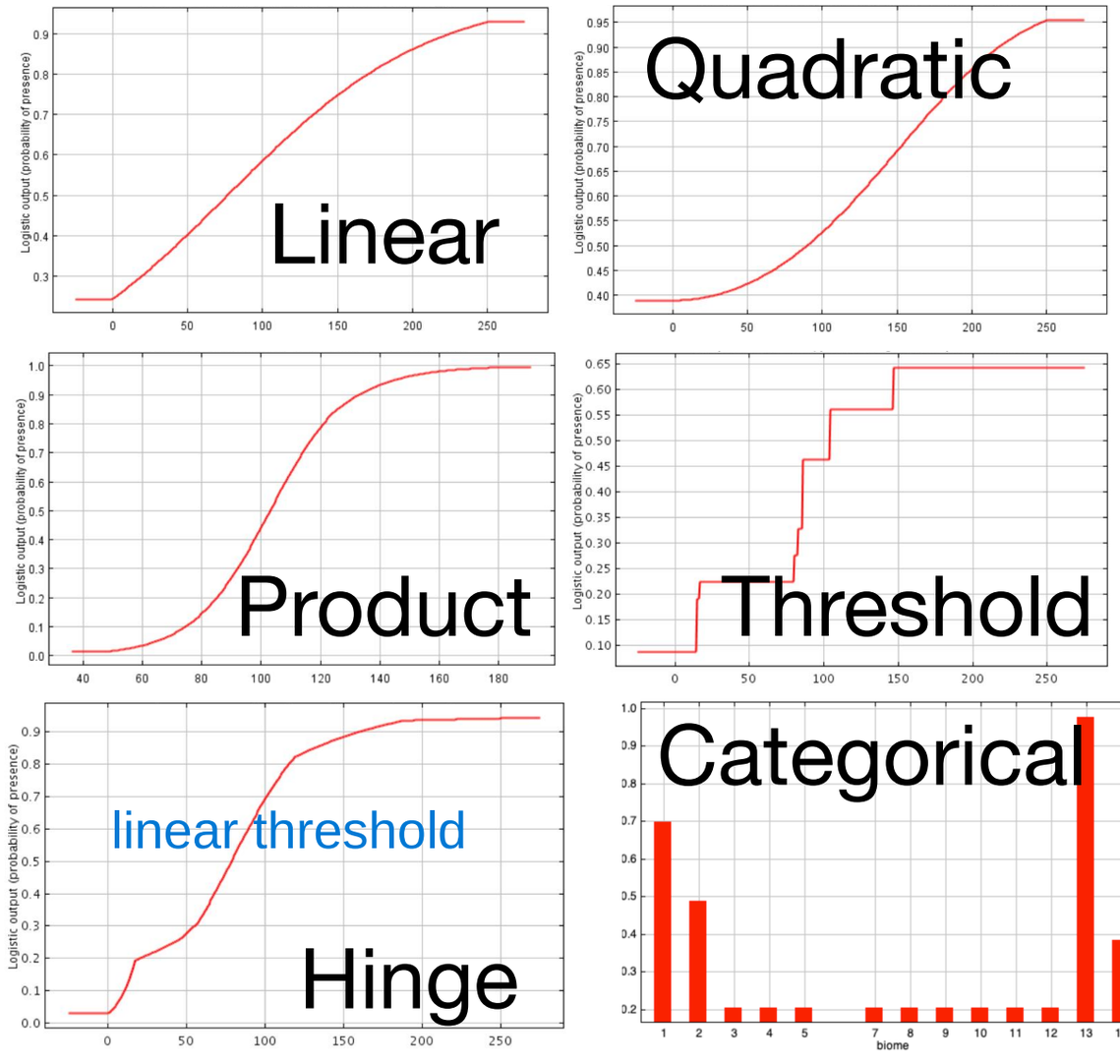
Cory Merow, Matthew J. Smith and John A. Silander, Jr

paper goes into more technical detail



Maximum Entropy (MaxEnt)

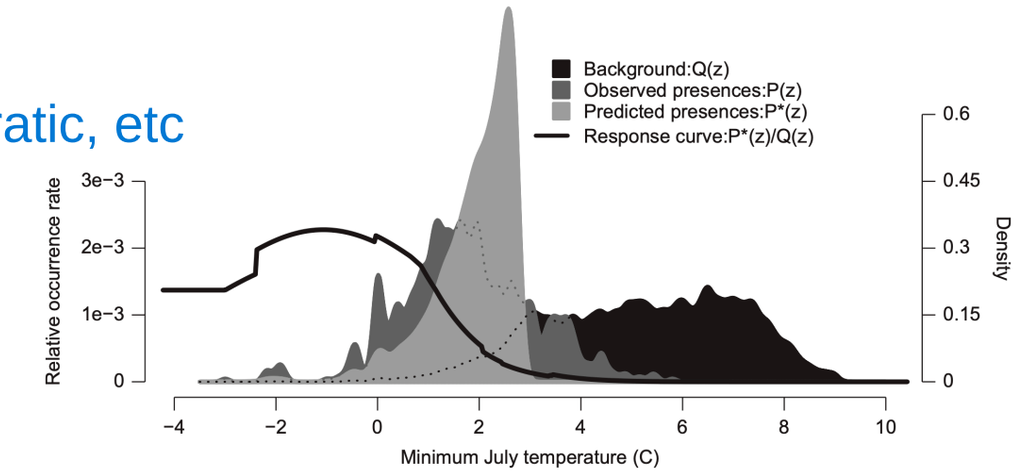
- Features = nonlinear functions of the environmental variables
  - Linear (mean)
  - Quadratic (variance)
  - Product (covariance, interactions)
  - Threshold (step function)
  - Hinge (“linear threshold”)
  - Categorical



can combine all of these nonlinear functions of predictors and you have to ask yourself whether that's a biologically relevant curve or not

Maximum Entropy (MaxEnt)

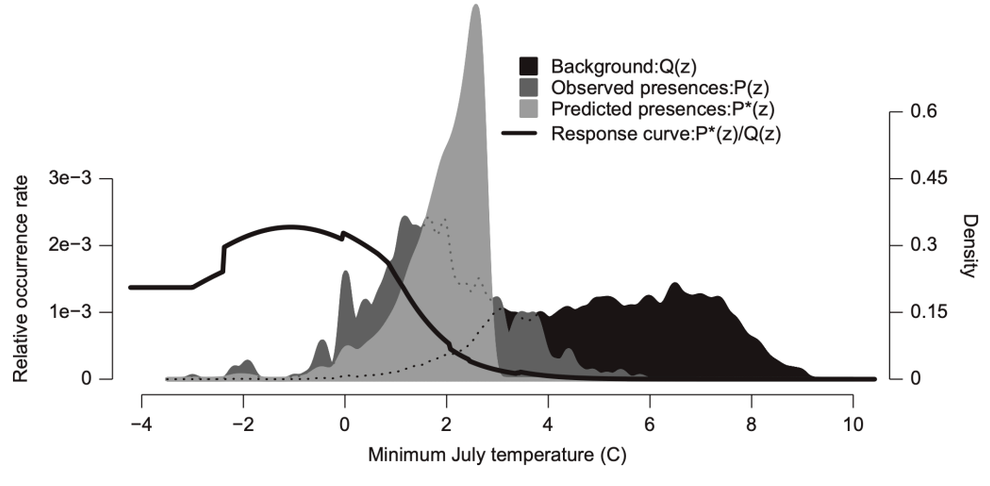
- Six key decisions
  - Where to select **background** samples (and how many)
  - Which **feature types** to include? linear, quadratic, etc
  - **Regularization**: Controls model complexity
  - How to deal with **sampling bias** if present
  - Type of **output**
  - Model **evaluation** just briefly touch on this here

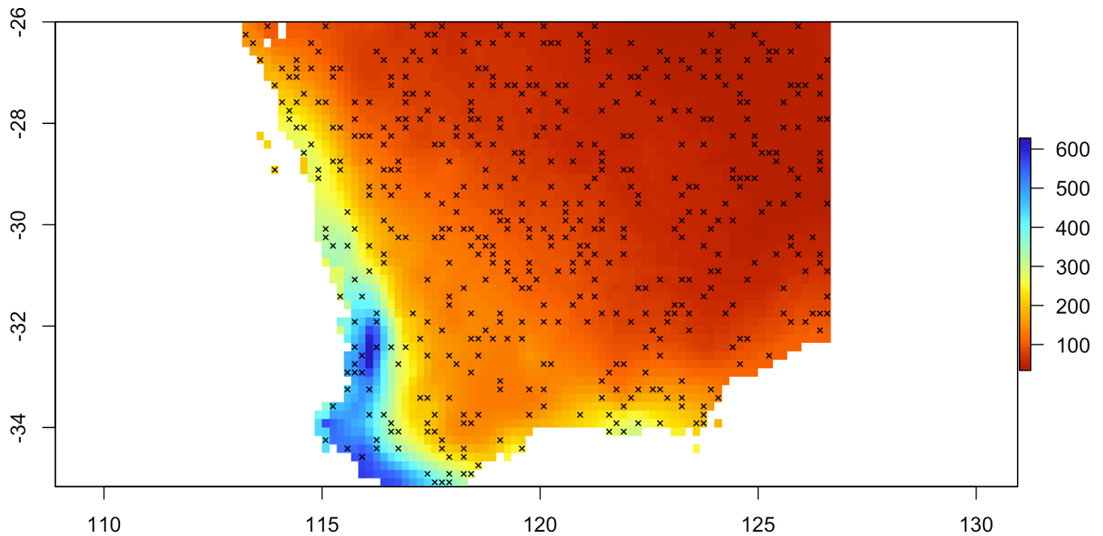


Background samples

- Region from which background samples are selected will influence maxent predictions

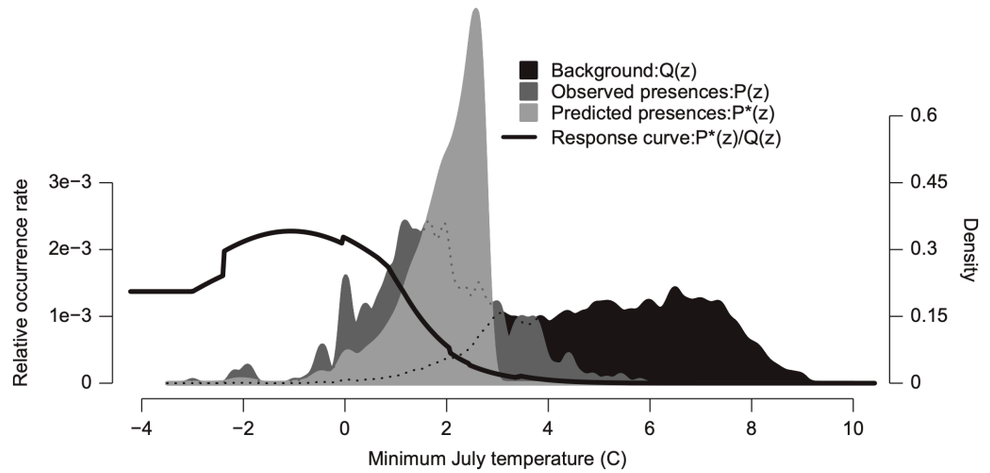
- Important consideration when extrapolating to new environments



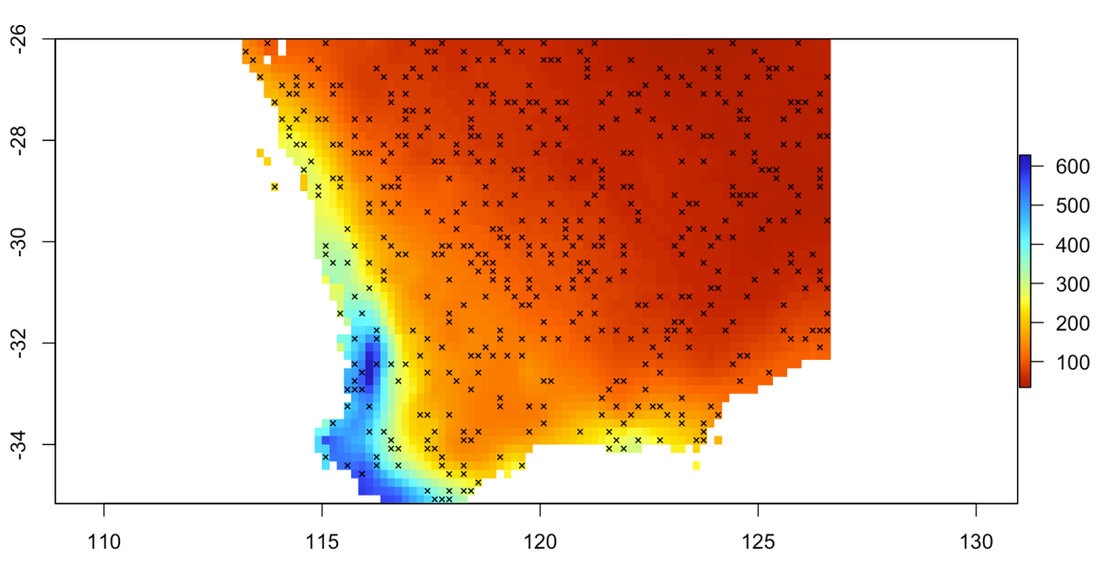


Background samples

- How to select:
  - What environmental conditions do you want to contrast against species presence?
  - For reserve selection?
    - Within the species current range
  - For invasions globally?
    - From within biomes in which the species could occur
  - Most relevant are areas species could reach via natural dispersal



can also use all the ecoregions in which the species occurs and select all of the ecoregions adjacent to those

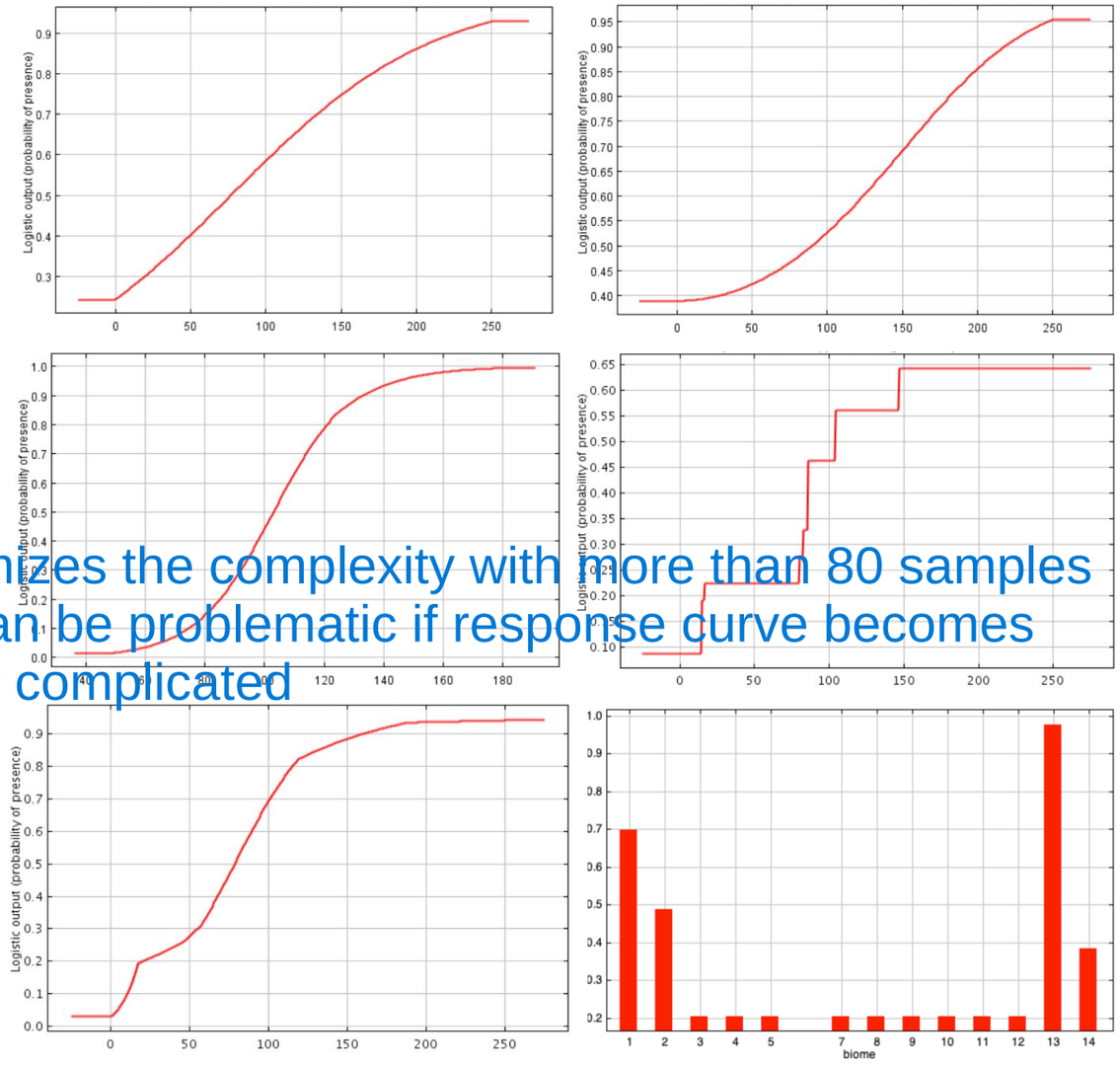


if you want a model that's very tight to known occurrences, then you probably want to restrict your background to where the species occurs -- ie. don't consider unrealistic environments against realistic ones

contrast: invasive species consider all background samples from all mediterranean biomes locally so that you get a very general model -- select based on dispersal ability of species

Features

- Can be highly complex or simple
- By default, determined by the number of occurrence records (> 80, all features are used)
- Can result in unrealistic, overly complex response curves that model noise rather than signal
- Hard to interpret



maximizes the complexity with more than 80 samples
this can be problematic if response curve becomes overly complicated

Features

- How to select:
  - If predictive accuracy is the only goal:
    - Let the machine learning algorithm work its magic
  - If understanding and/or projecting is the goal:
    - Minimize correlations among variables
    - Identify what you think are appropriate response curves in the conceptualization stage
    - Goal is to produce a parsimonious, interpretable model

the algorithm isn't sensitive to correlations but it's more difficult to determine variable importance if highly correlated

go for simplicity if you want to interpret

Regularization

- Regularization coefficient = beta

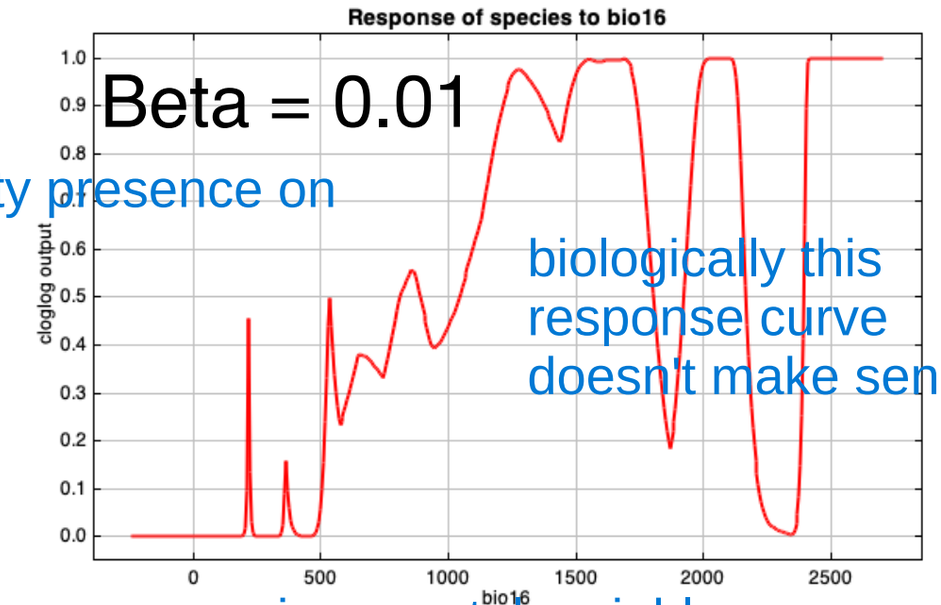
- Maxent selects feature complexity using beta

- constrains model complexity and limits overfitting

can try a series of beta values to see where the model seems to have reasonable complexity

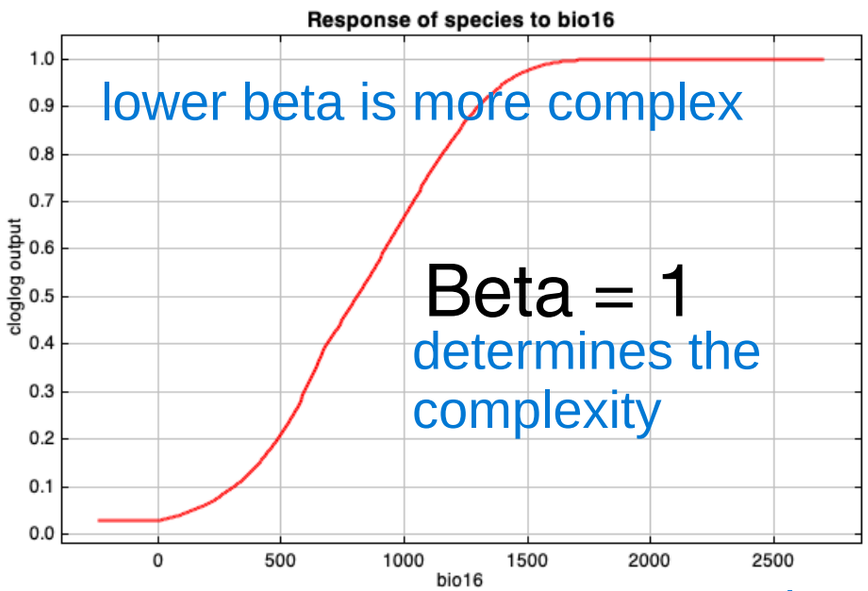
modeling noise in the graph below

probability presence on y axis



biologically this response curve doesn't make sense

environmental variables



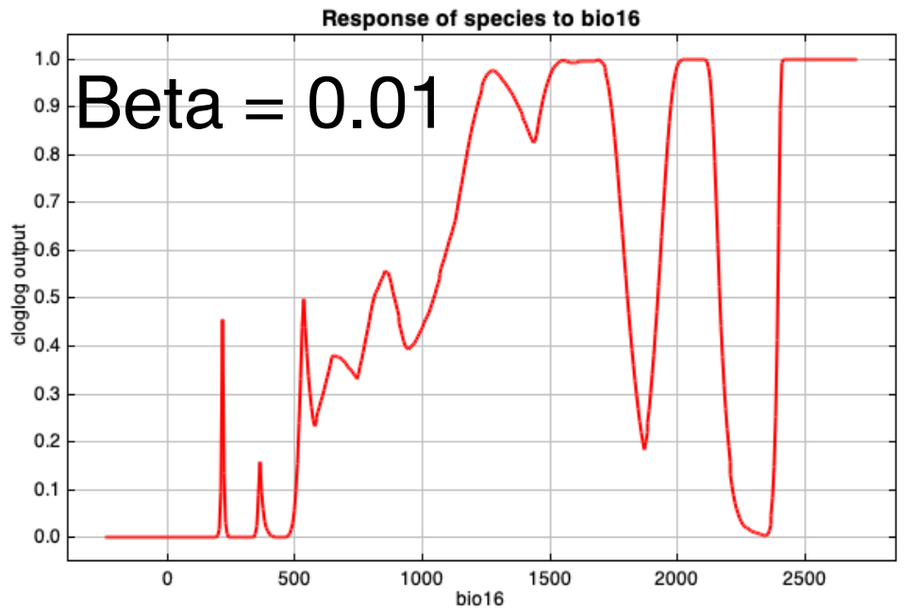
determines the complexity

Maxent tunes Beta parameter using training and test data

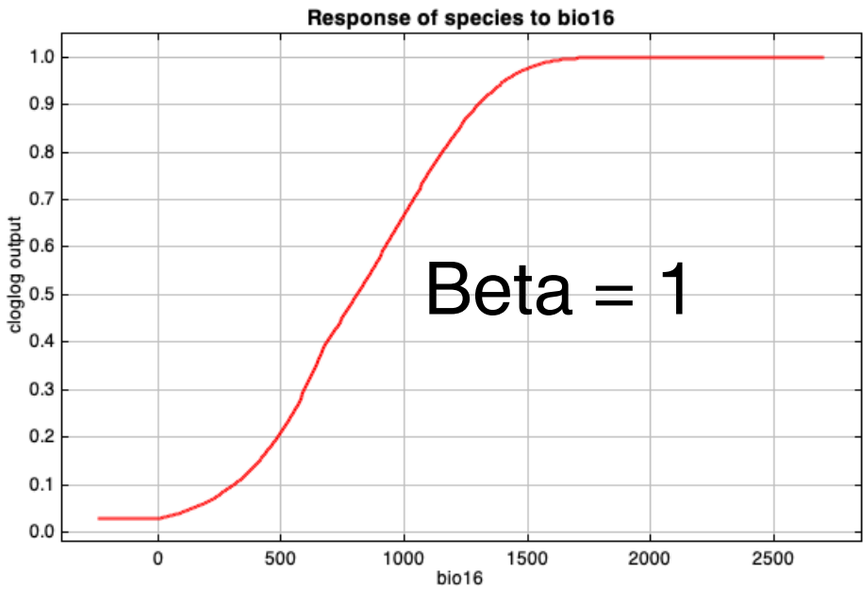
Regularization

- How to select:
  - Explore a range of values and select a value that:
    - maximizes model fit on a cross-validation set
    - Is parsimonious (AIC)

      use AIC outside the MAXENT to get the most parsimonious in MAXENT



Beta cannot be 0



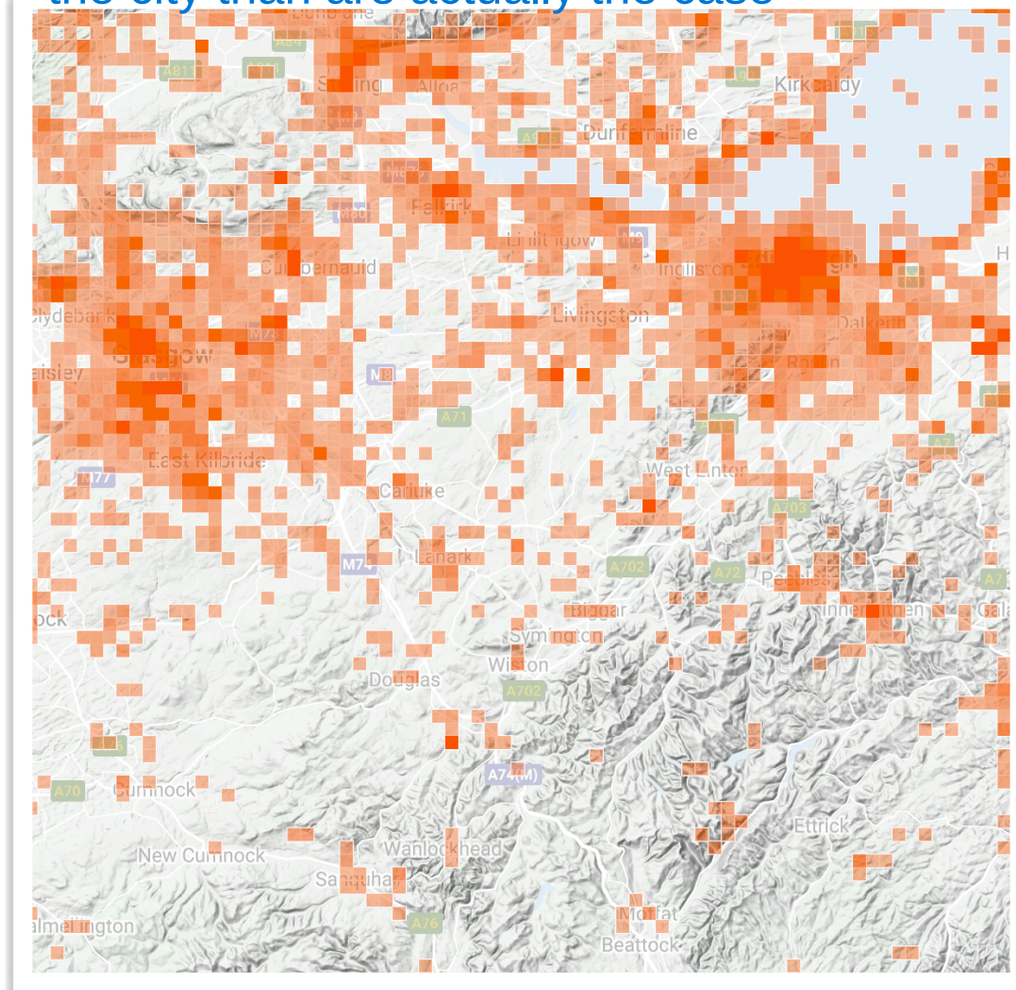
Sampling bias

- Default assumption is that all locations are equally likely to be sampled

- Preferential sampling: stone locations are more visited than others (roadsides, urban areas)

- Cannot distinguish habitat selection from observation bias

preferentially sampled locations by the city caused by more people looking in the cities than in rural areas
--this is likely making higher presence areas in the city than are actually the case

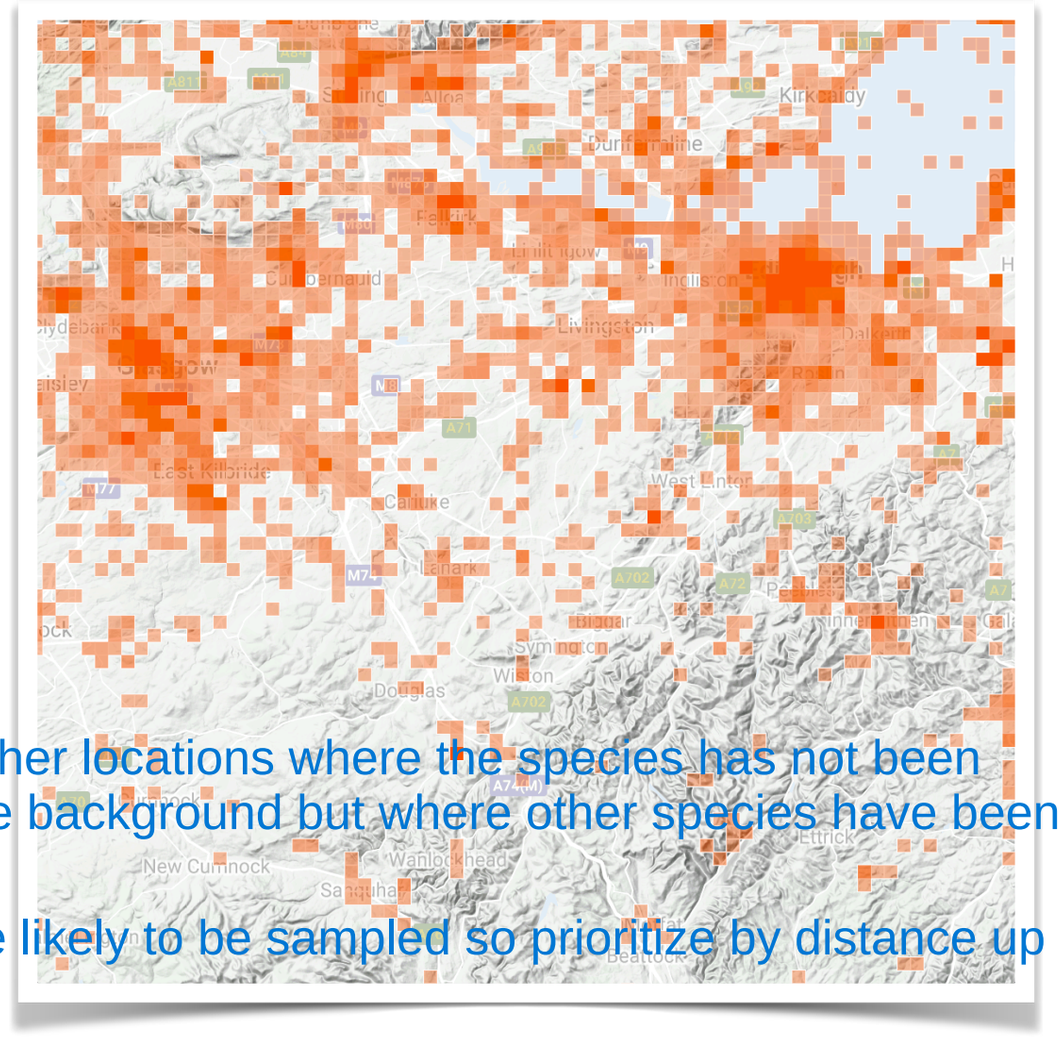


Sampling bias

use kriging to make a density surface that reflects the background and then that allows the same bias to be in the background, essentially canceling the bias out

- How to address sampling bias
  - Create a biased background selection
    - Modifies selection of background points so they are more likely to be sampled where presence records are dense
    - Cancels out bias because it is equal across presences and background points
    - Target Group Sampling (TGS)

use all of the other locations where the species has not been observed as the background but where other species have been observed



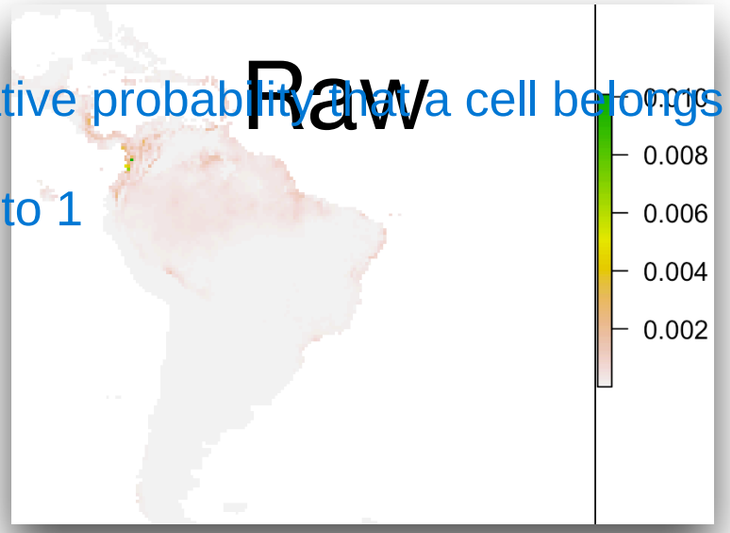
- Use a biased prior

ie. lower elevations are more likely to be sampled so prioritize by distance up the mountain

  - Distance to roads, urban centers, elevation

Output format

the raw output that comes out of MAXENT is the relative occurrence rate (relative probability that a cell belongs to a presence location)
output is very small because the output of the grid of cells (pixels) has to sum to 1



- Raw

  cumulative output takes each pixel and sums the output of pixels with the same or less value. Then sums all that up and assigns the sum to that pixel and then does that for all pixels and converts to a percent scale.

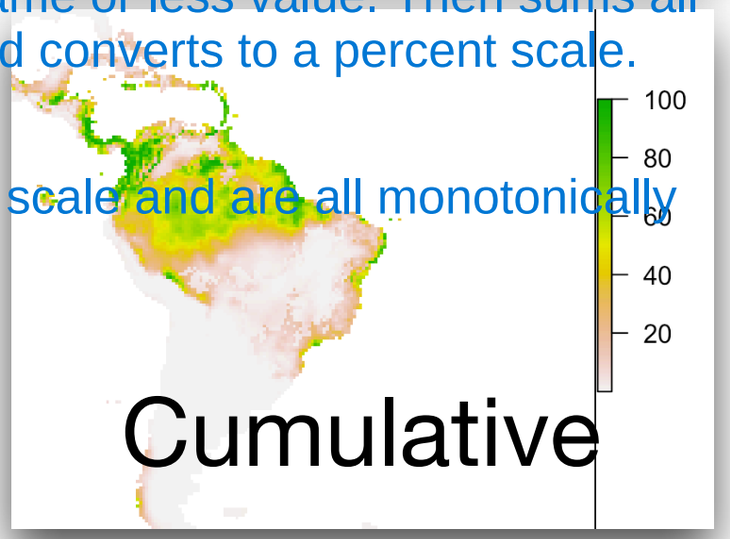
- Cumulative

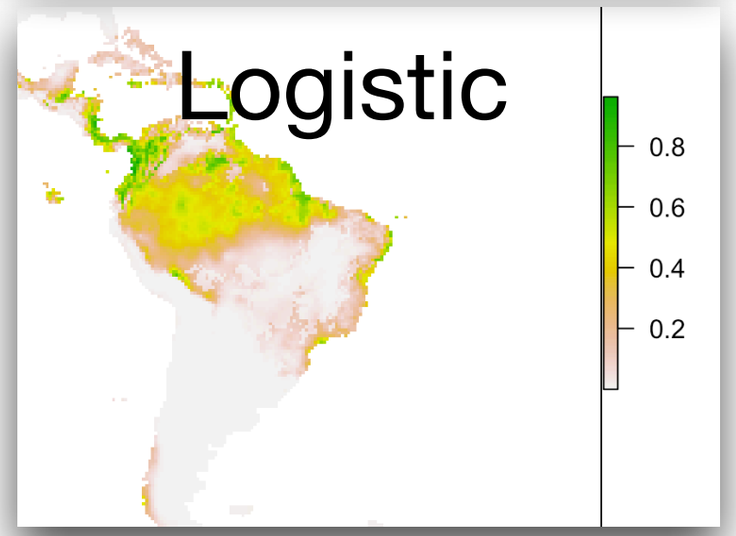
  if you take 10 pixels for raw, cumulative, and logistic they all have the same scale and are all monotonically related. It's just that the values of the data are re-scaled differently.

- Cloglog

- Logistic / Cloglog

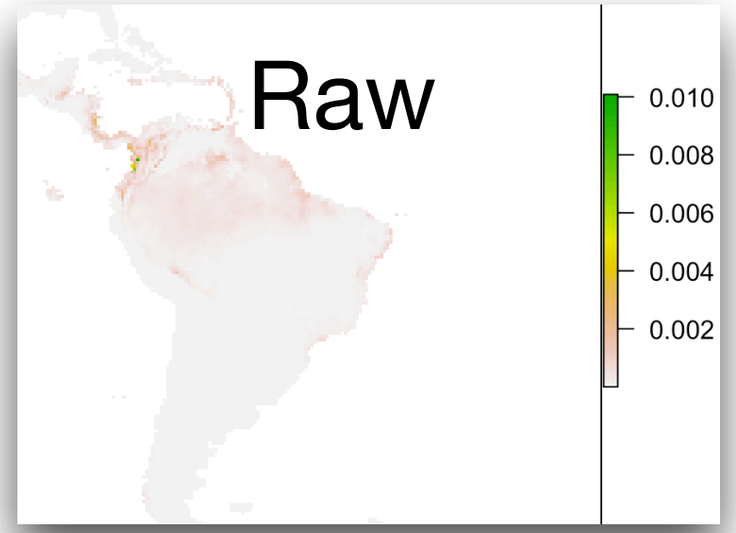
- Monotonically related, but scaled differently

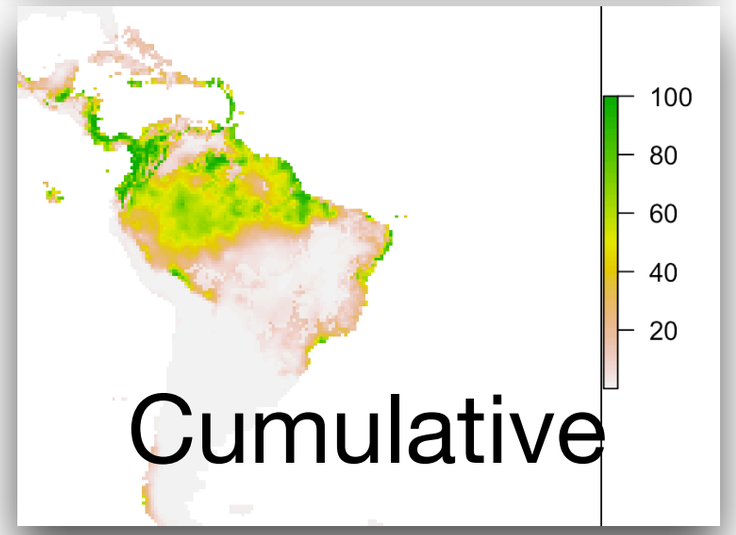


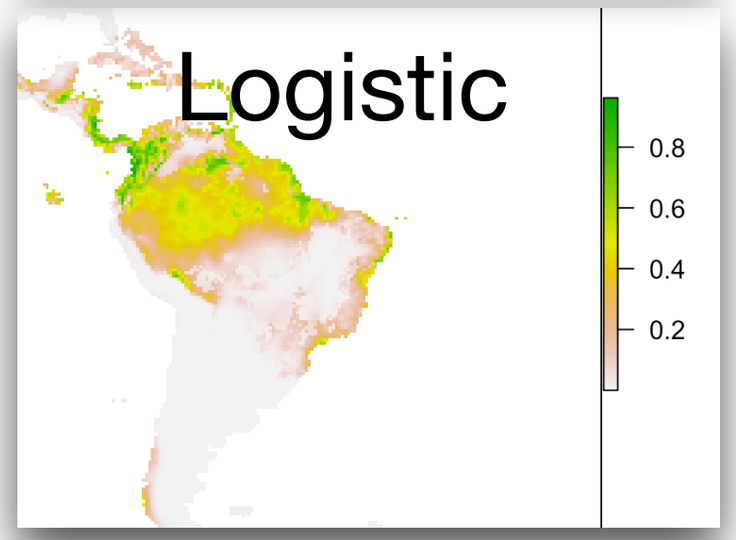


Output format

- Raw
  - This is the ‘raw’ prediction from Maxent’s exponential model
  - 0-1
  - Probability in each grid cell
  - **Sums to one over entire map**





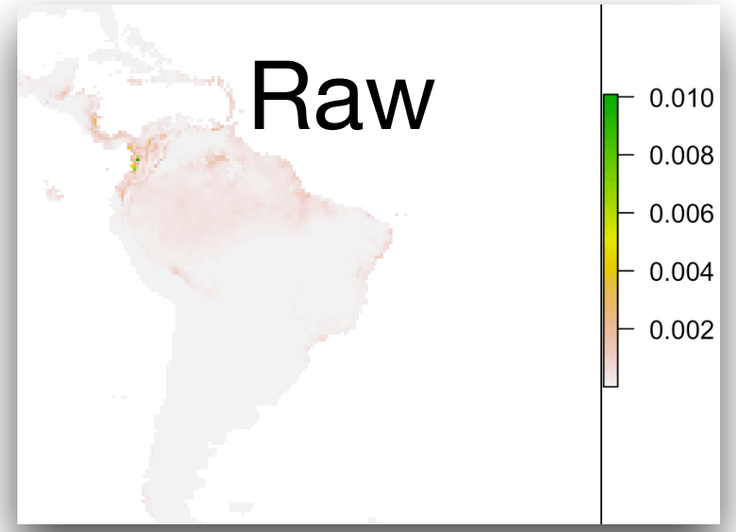


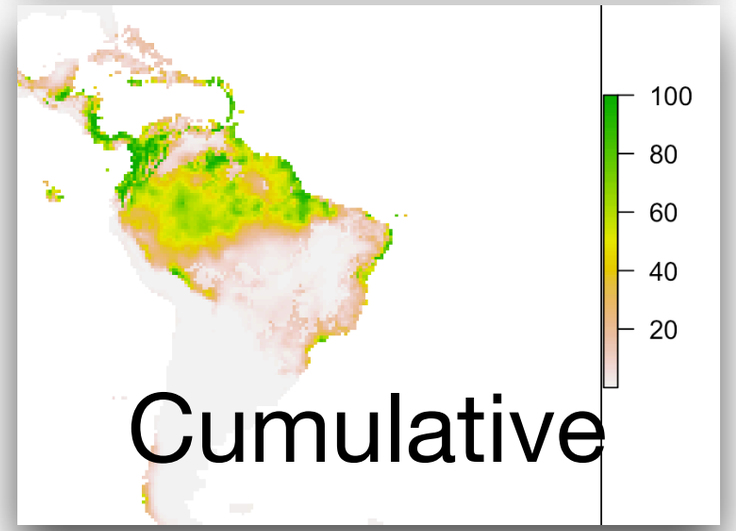
Output format

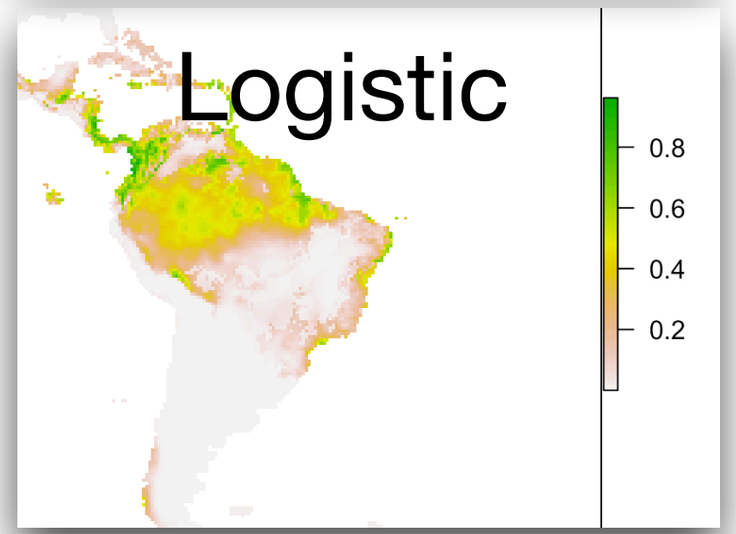
- Cumulative

  highest value overall represents the grid cell with the best conditions

  - 0-100
  - Value at a grid cell = sum of the probabilities of all grid cells with probability < the grid cell x 100
  - cumulative value = 100 for grid cell predicted as having the best conditions for the species (because all cells in the map would be <= this value)

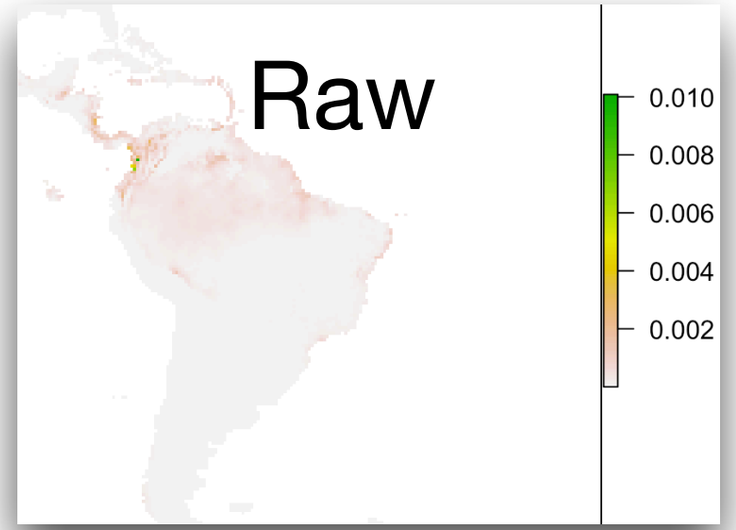


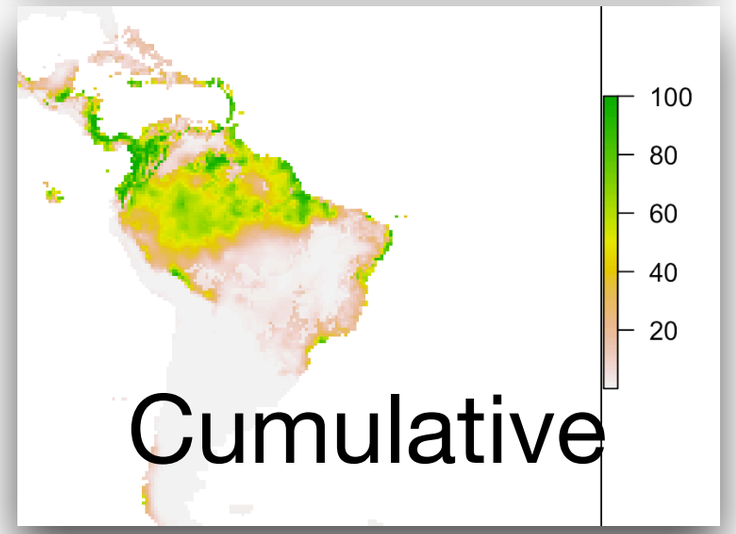


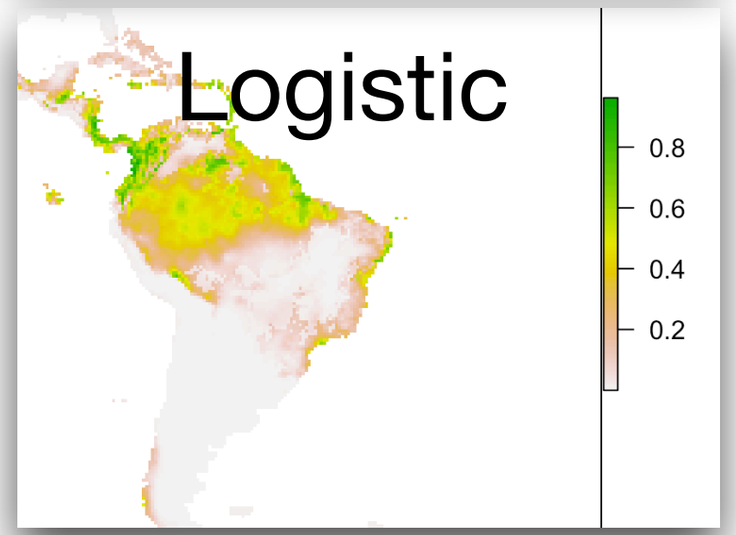


Output format

- Logistic / Cloglog similar techniques but not the same
  - Transformation of raw output into something more like probability of occurrence (the relative occurrence rate)
  - Estimated probability of presence in each grid cell, under strong assumptions: typical presences used for model training are from environmental conditions where probability of presence is around 0.5
  - Cloglog “estimates probability of presence assuming that the sampling design is such that typical presence localities have an expected abundance of one individual per quadrat, which results in a probability of presence of about 0.63.”

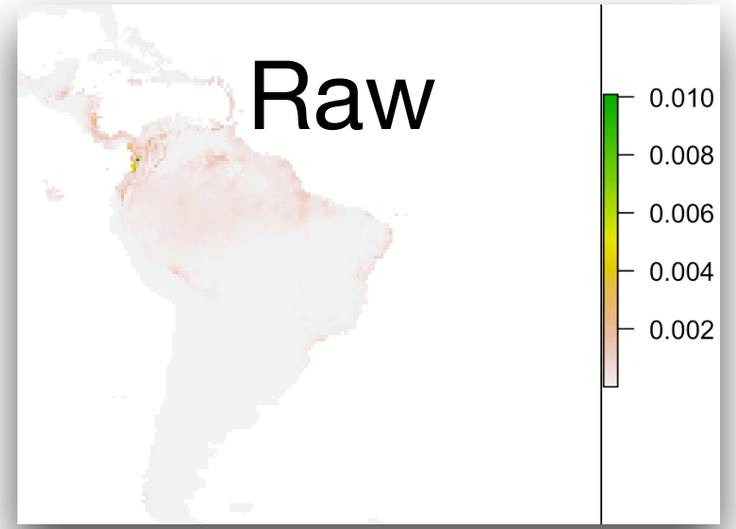


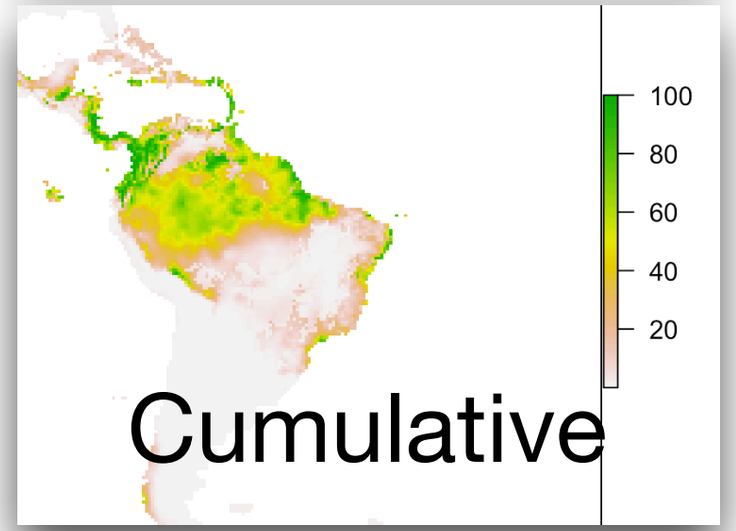


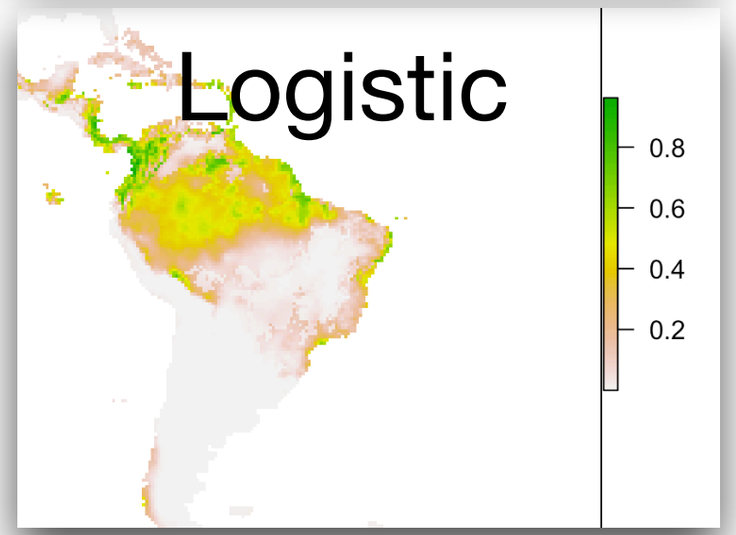


Output format

- How to select:
  - Use ‘raw’ when possible
    - Useful for comparing different models of the same species using different background samples, etc
  - Use ‘cumulative’ when predicted absence is of interest
    - Determining range boundaries
  - Avoid (not easy!) logistic given the strong assumptions and avoid CLOGLOG
    - Be careful when comparing models across species that differ in prevalence
    - Try to estimate prevalence if possible, or use a range of conservative values





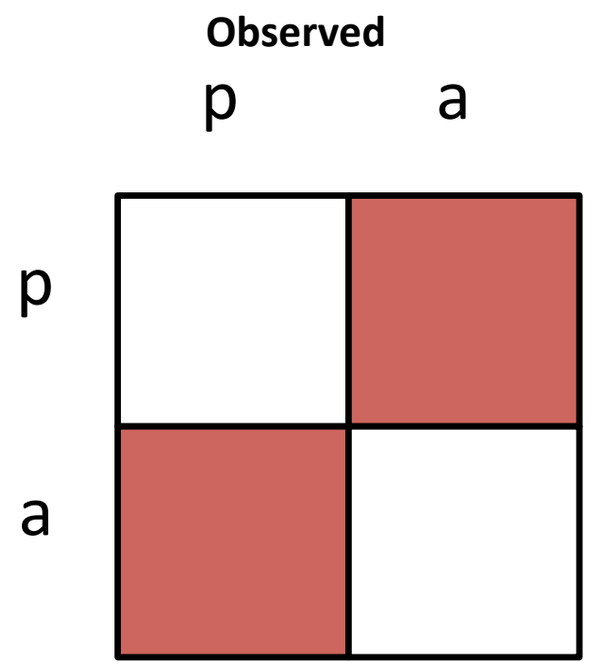


Maxent evaluation

- Use AIC to assess model fit (ENMTools)

AUC=Area Under the Curve

- AUC can be trouble because it is a P-A metric, not presence-background
  - Assess ability to distinguish presences from background locations, which could also be presences
  - Comparisons between models are only valid if but for the same landscape, background sample, and species using the same test data
  - Most appropriate for species near range equilibrium, sampling intensity is high, and background choice reflect biology

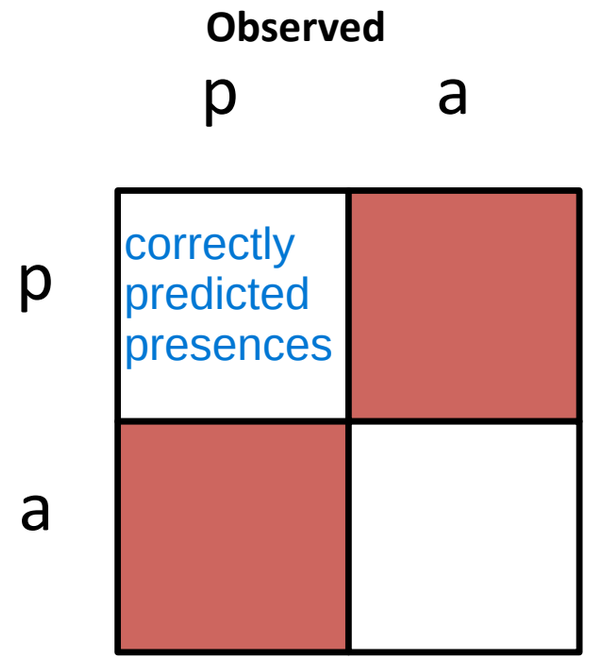


measures how well the model can discriminate presences from absences

Maxent evaluation

statistically independent: two datasets collected in different ways from the same area

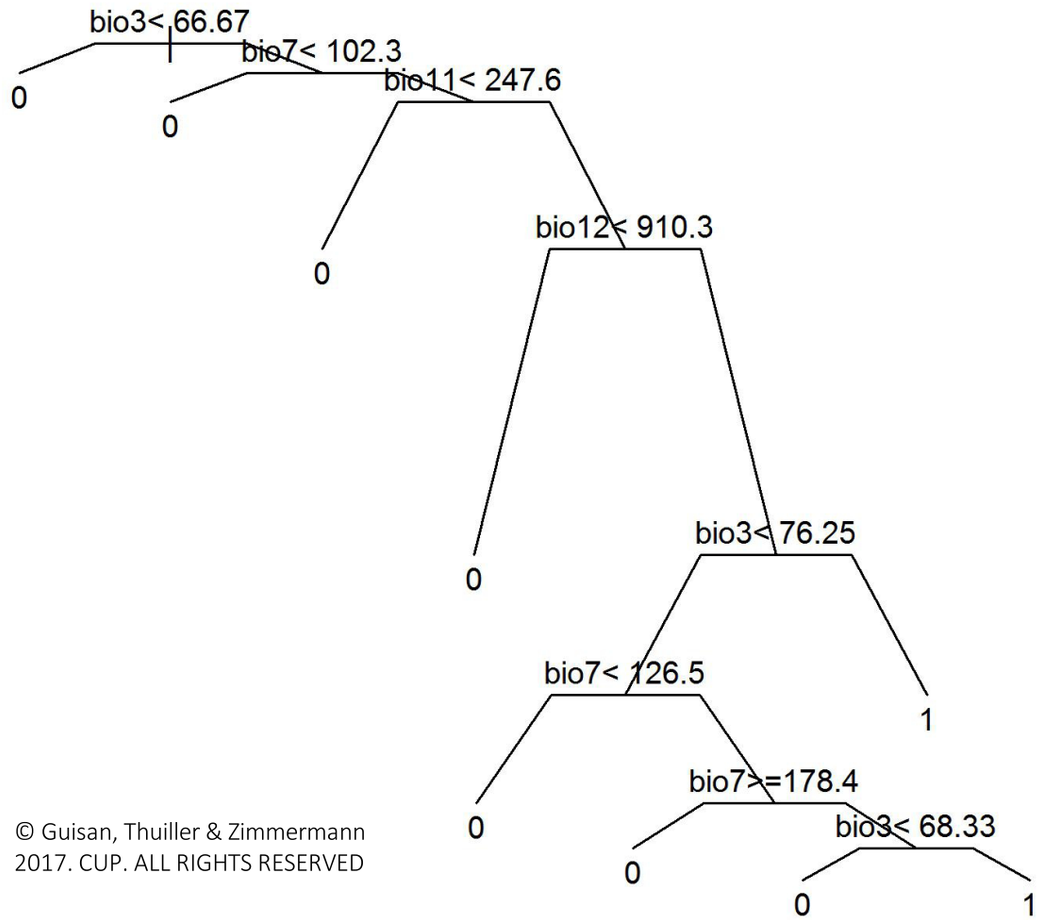
- How to select:
  - Where possible, use statistically independent, k-fold cross-validation
  - Scrutinize sensitivity = correctly predicted presences
  - Estimate uncertainty from k-folds if you have 100s of 1000s of data points the k-folds don't matter as much bc so much data
  - Avoid measures based on specificity = correctly predicted absences



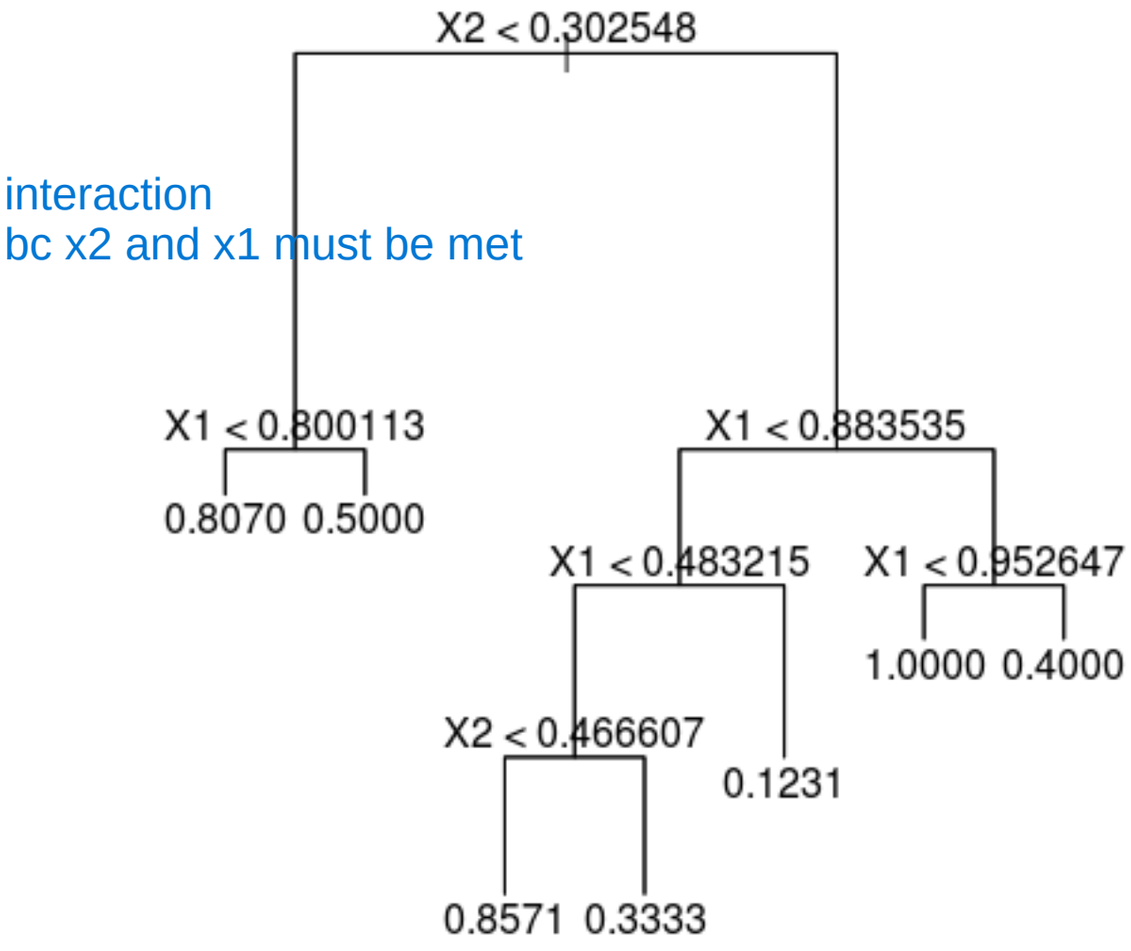
AUC includes 1-specificity which is why you cannot include specificity in the measures

Classification & regression trees

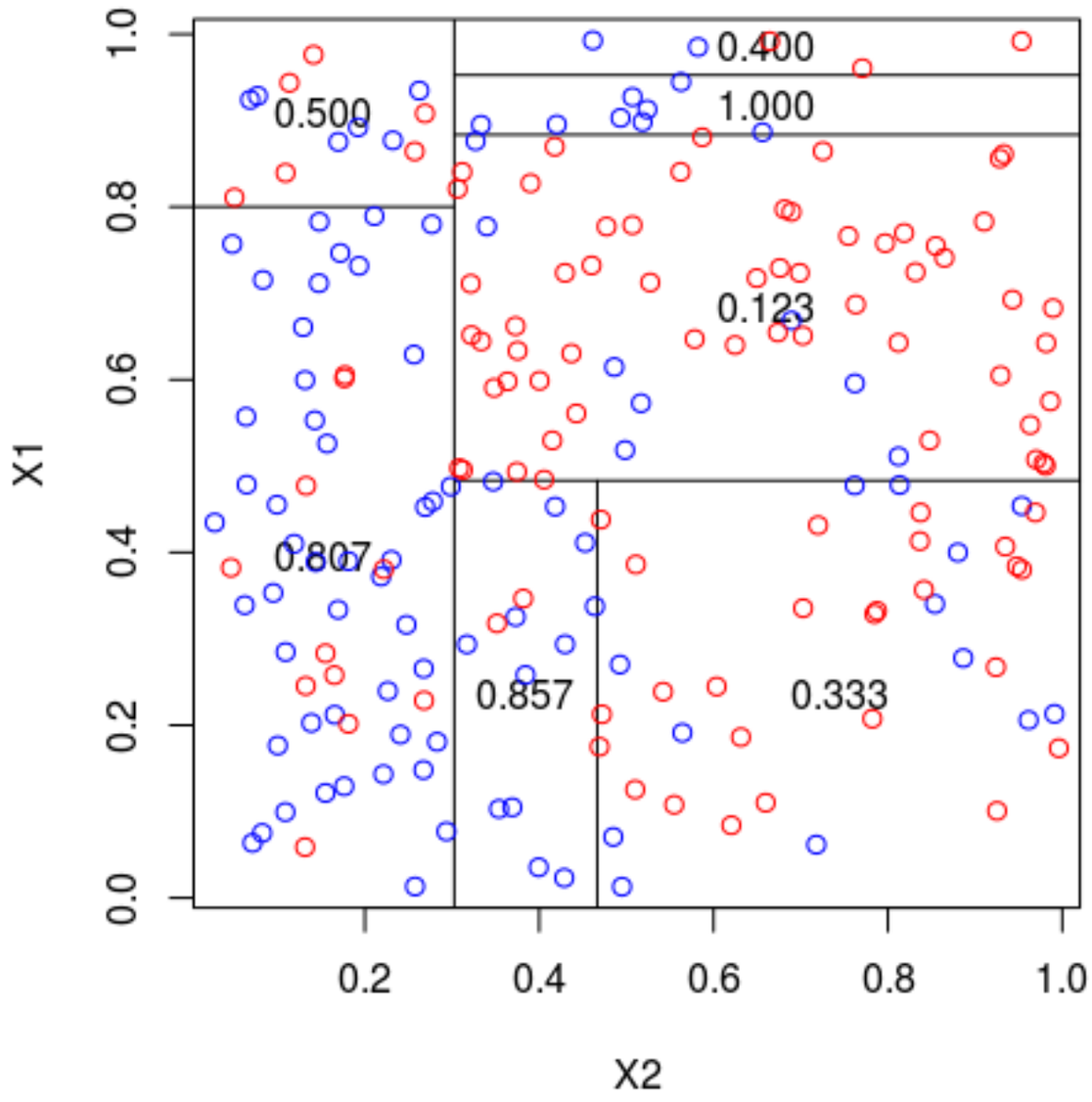
- Recursively splits the data
- Can be conceptualized as a series of if-then statements



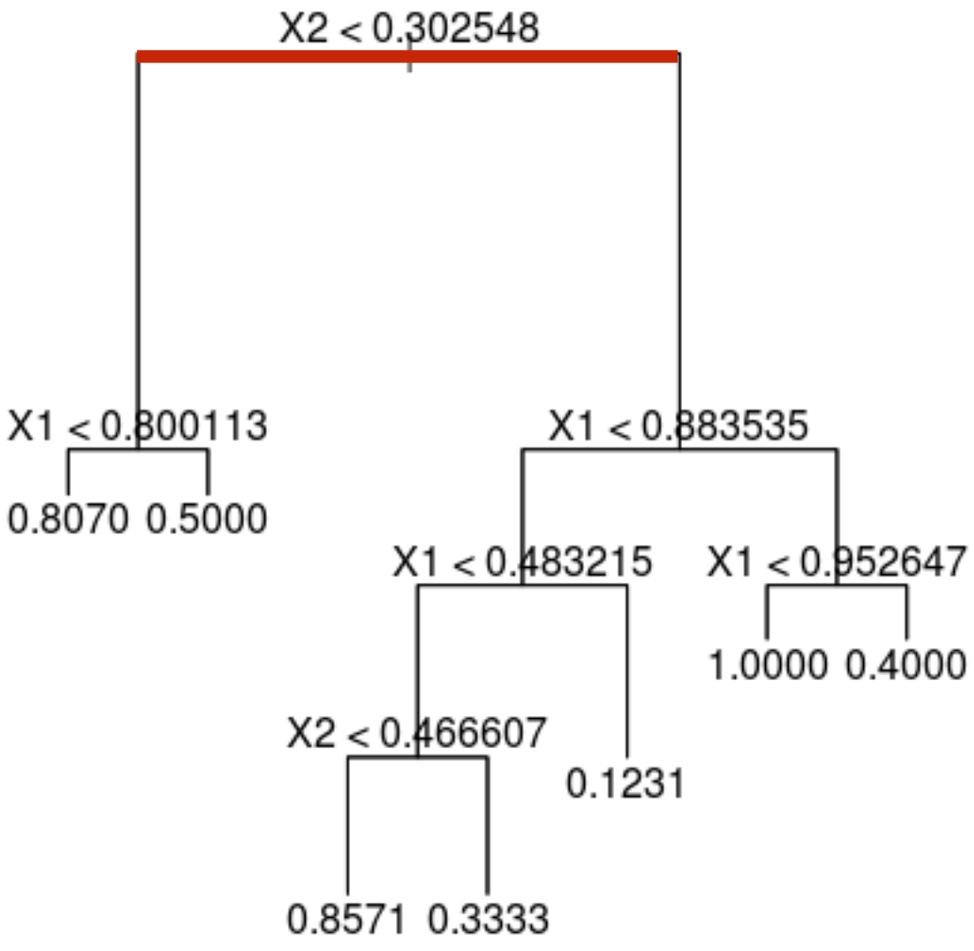
Classification & regression trees

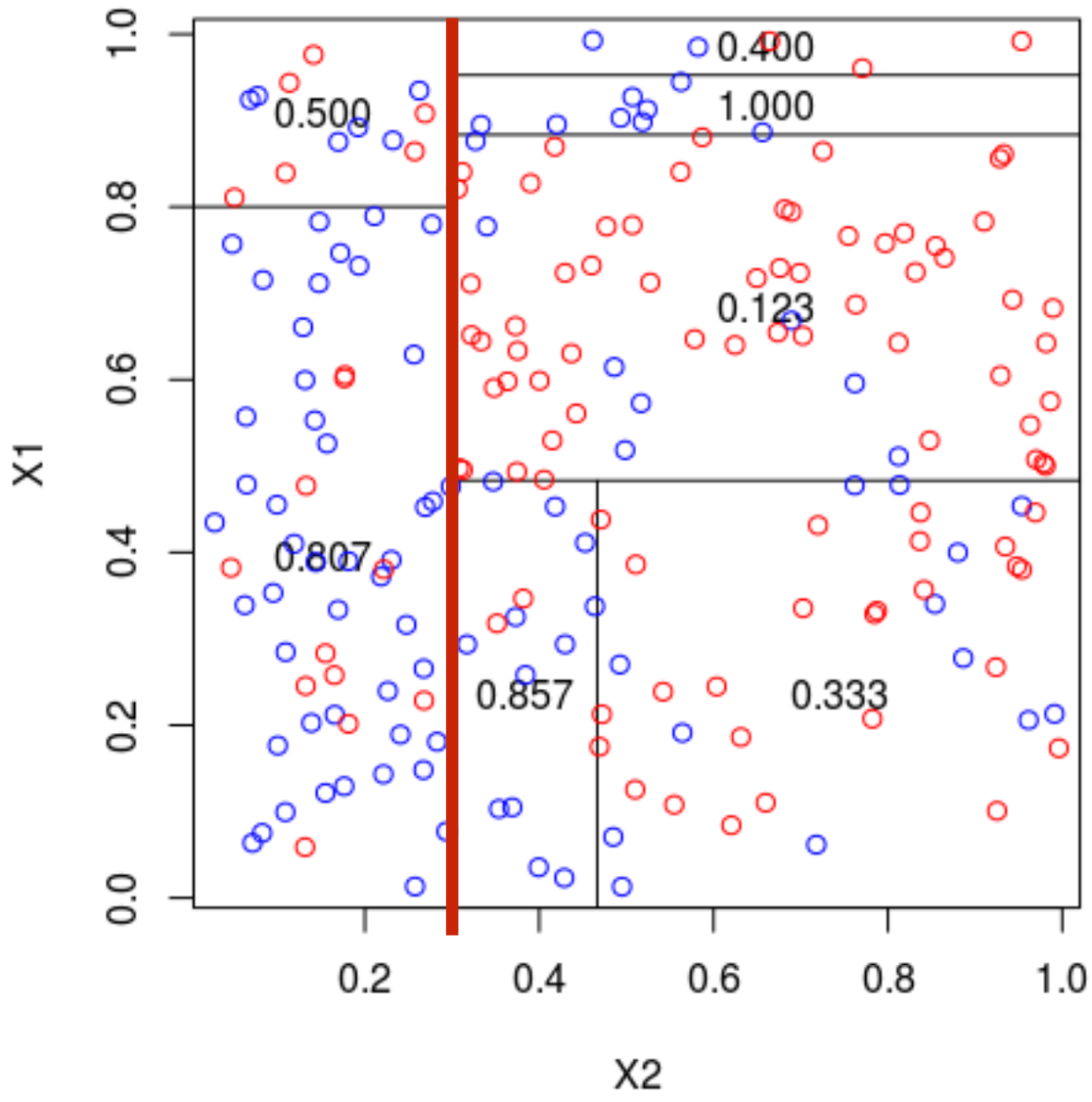


interaction
bc x2 and x1 must be met

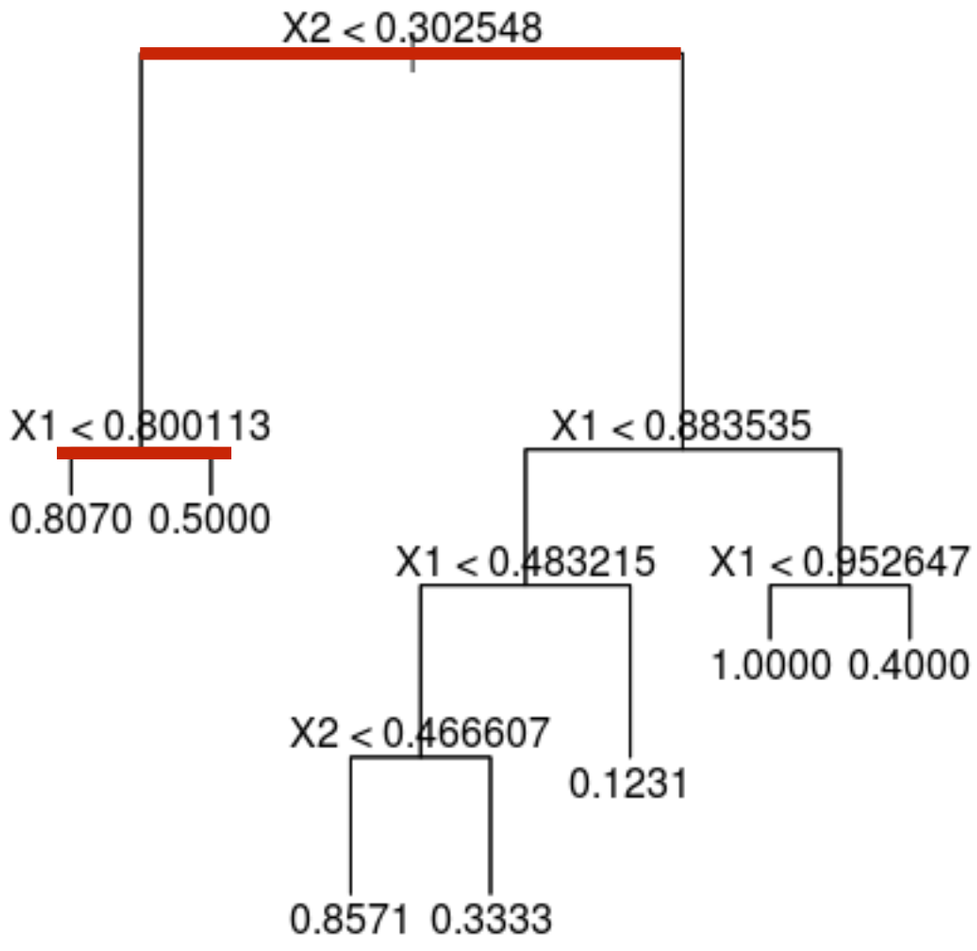


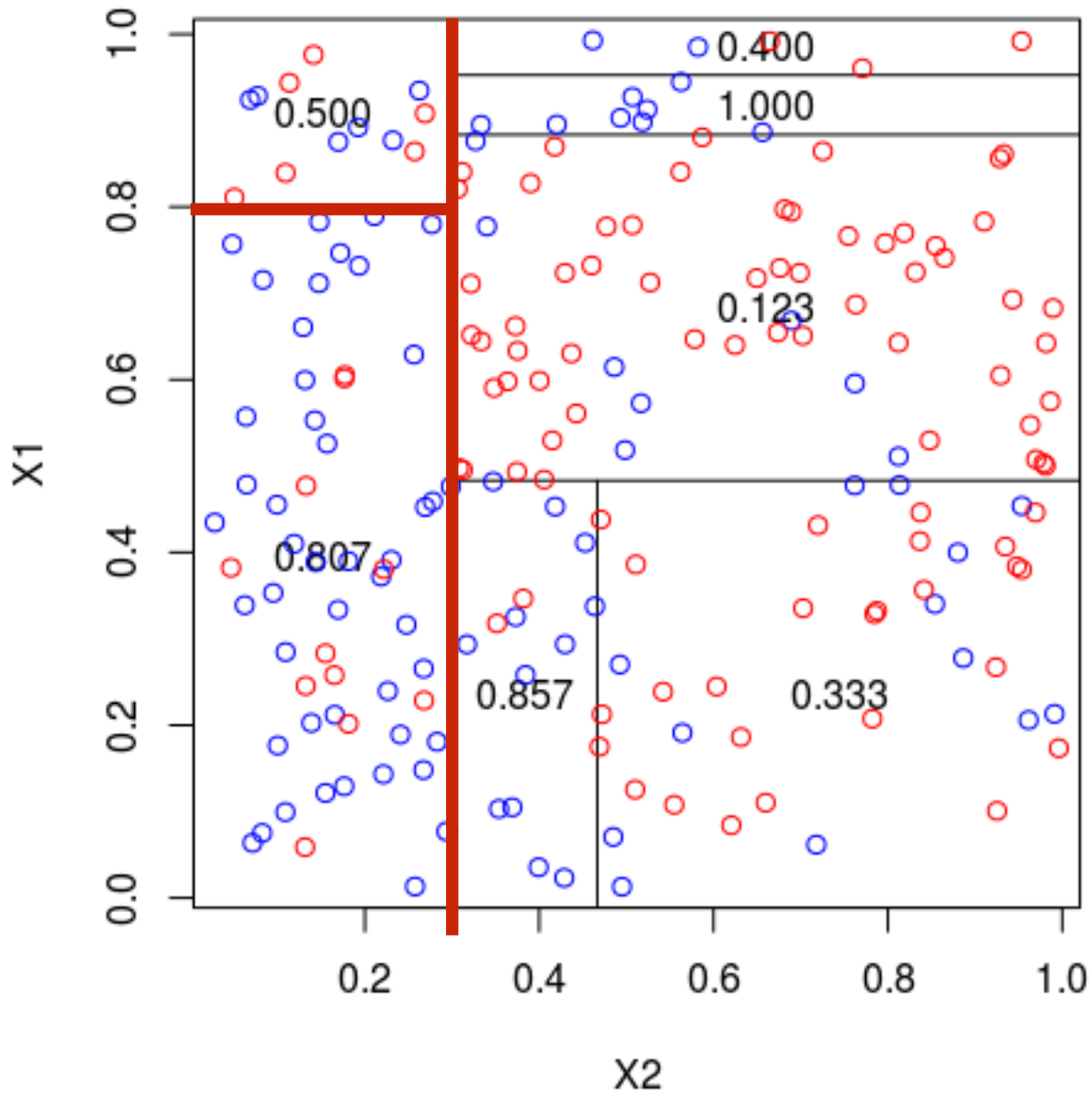
Classification & regression trees





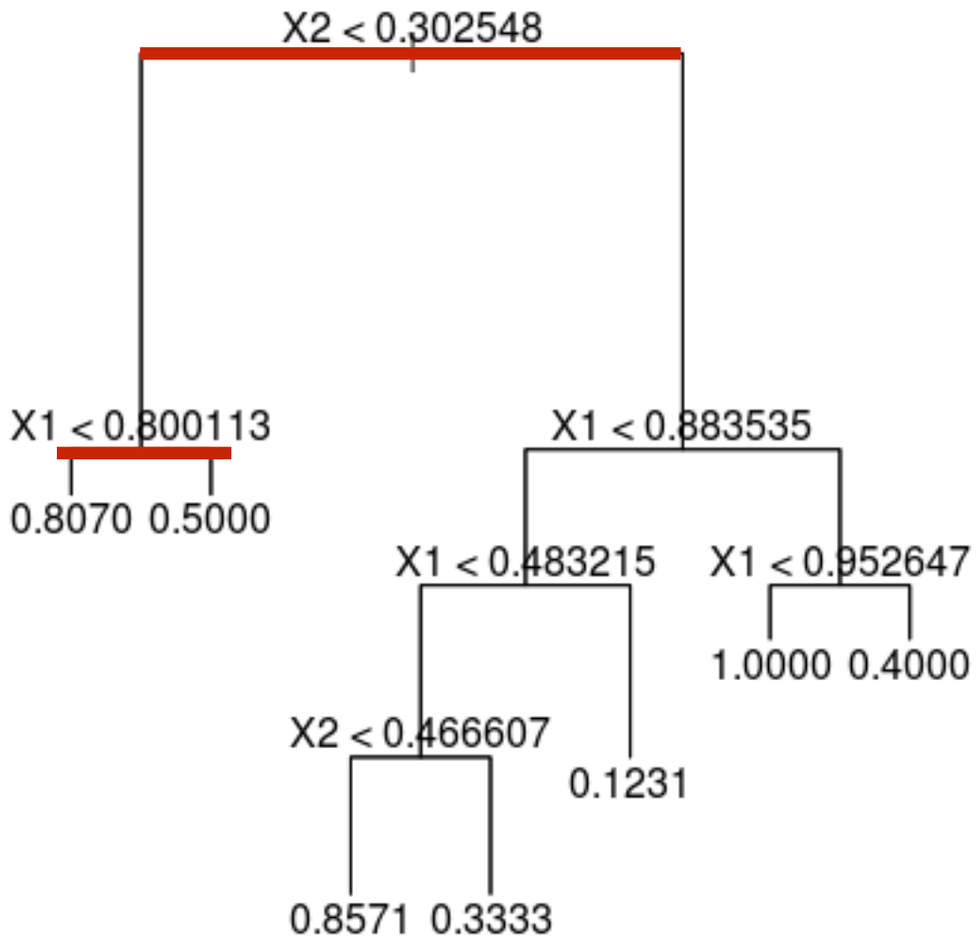
Classification & regression trees

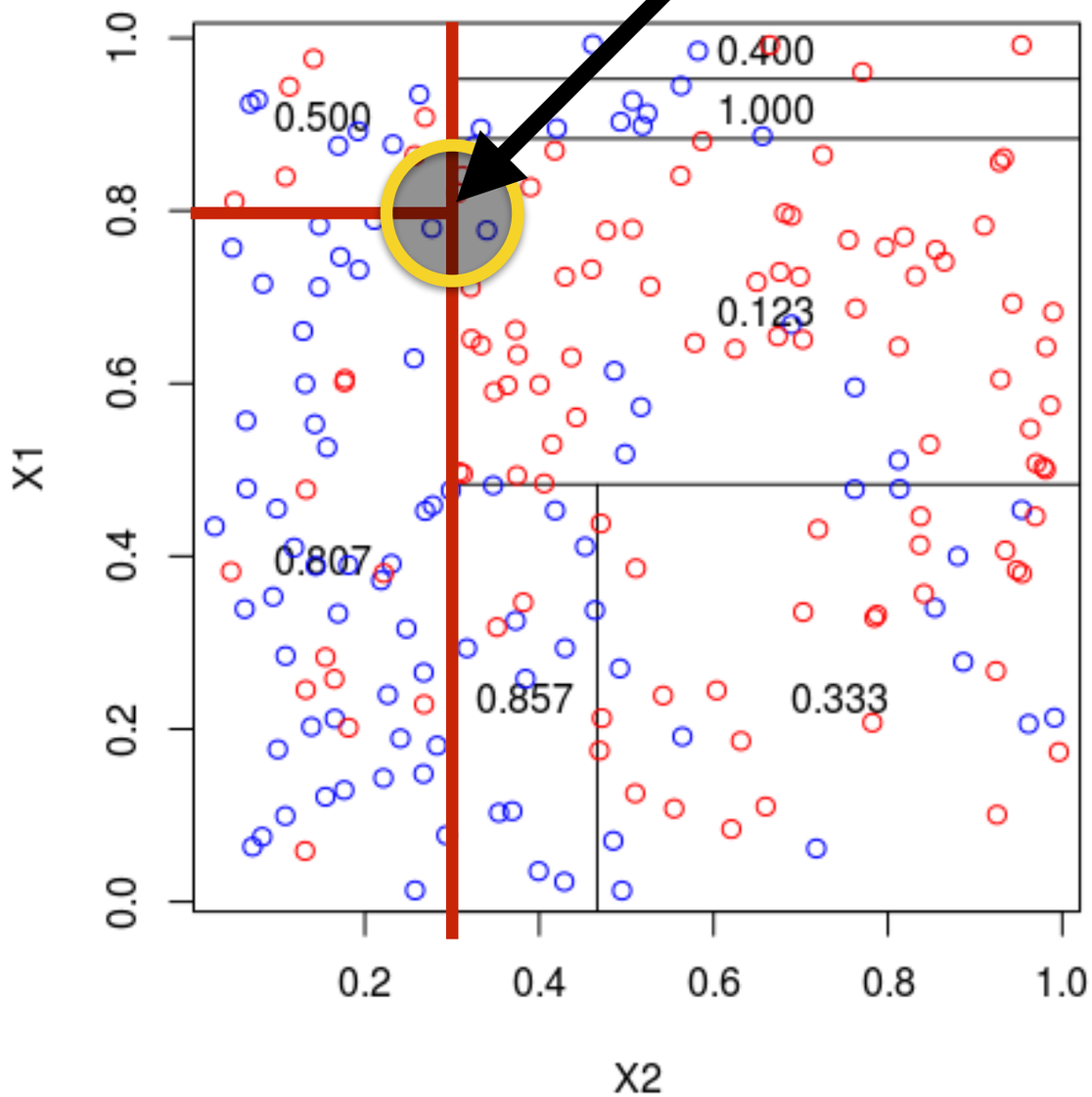




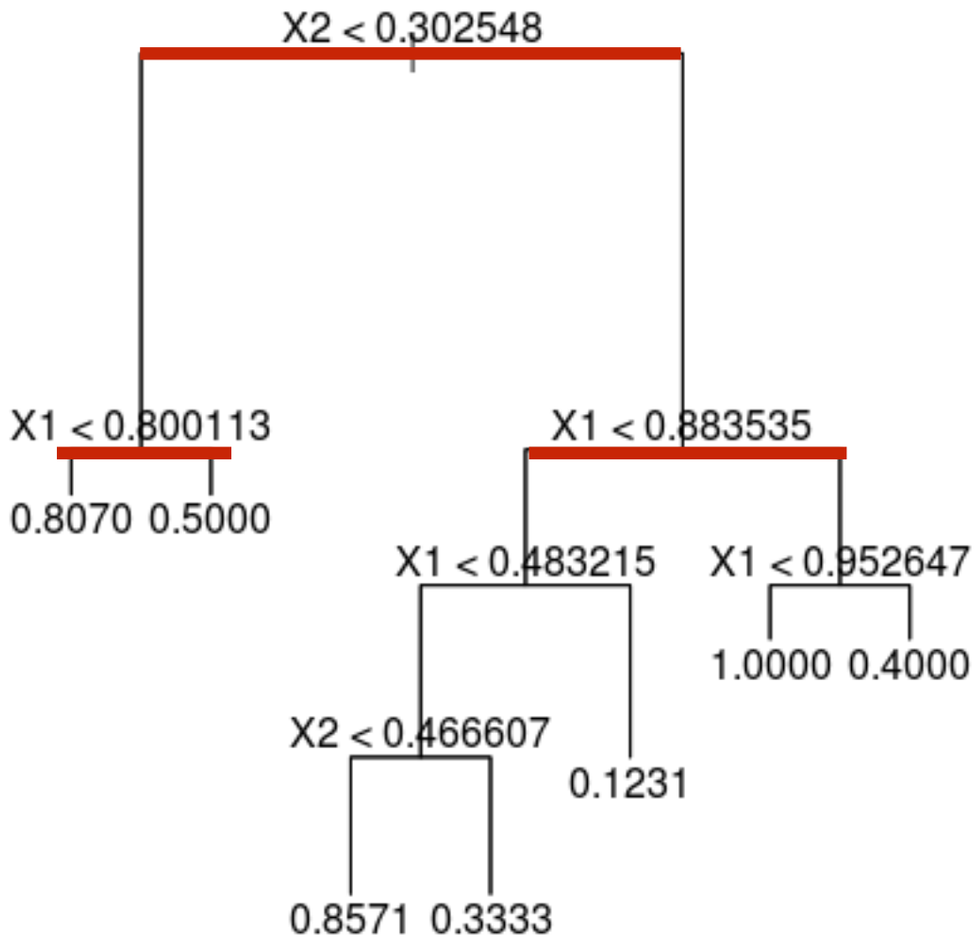
Classification & regression trees

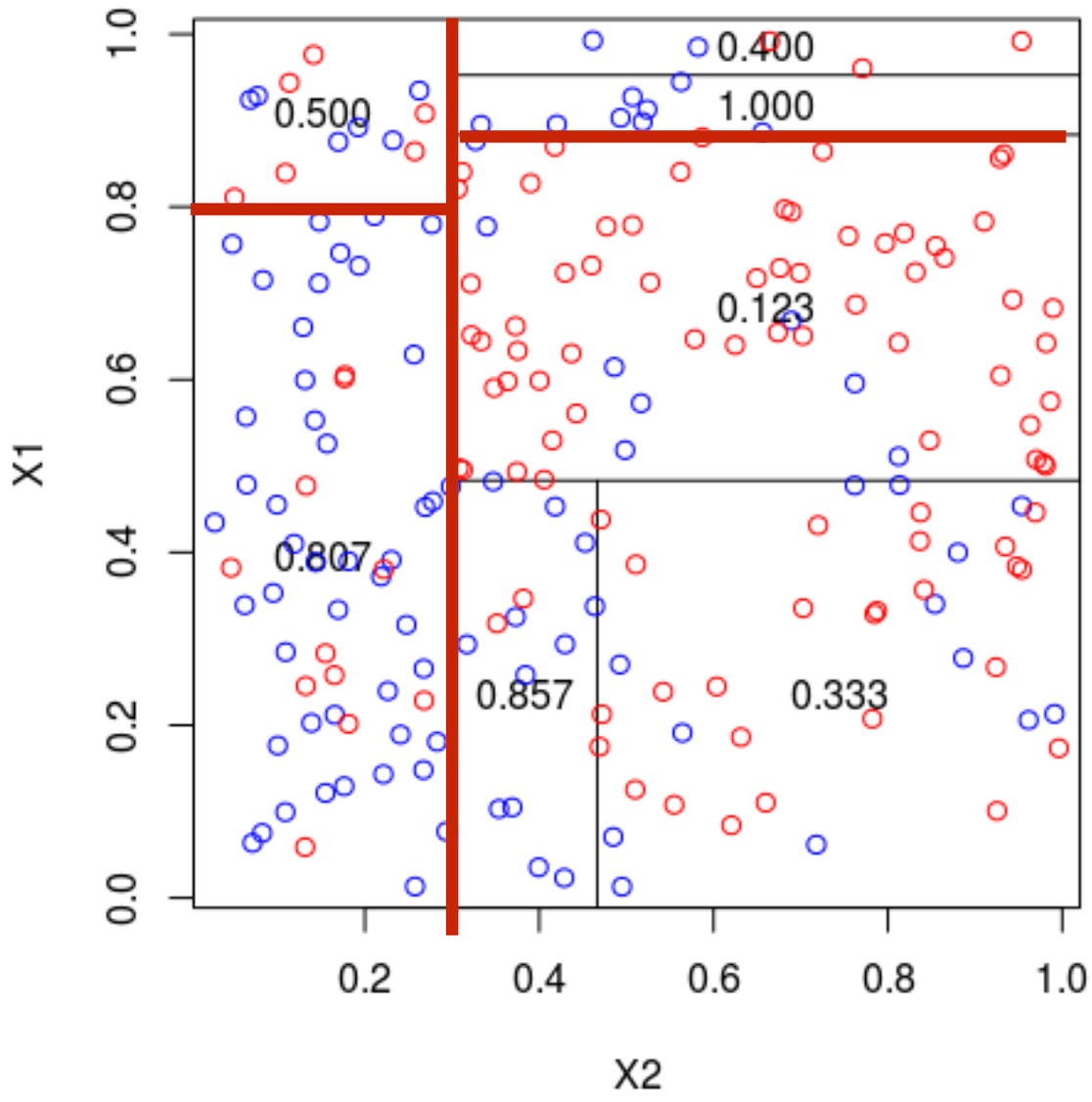
Interaction



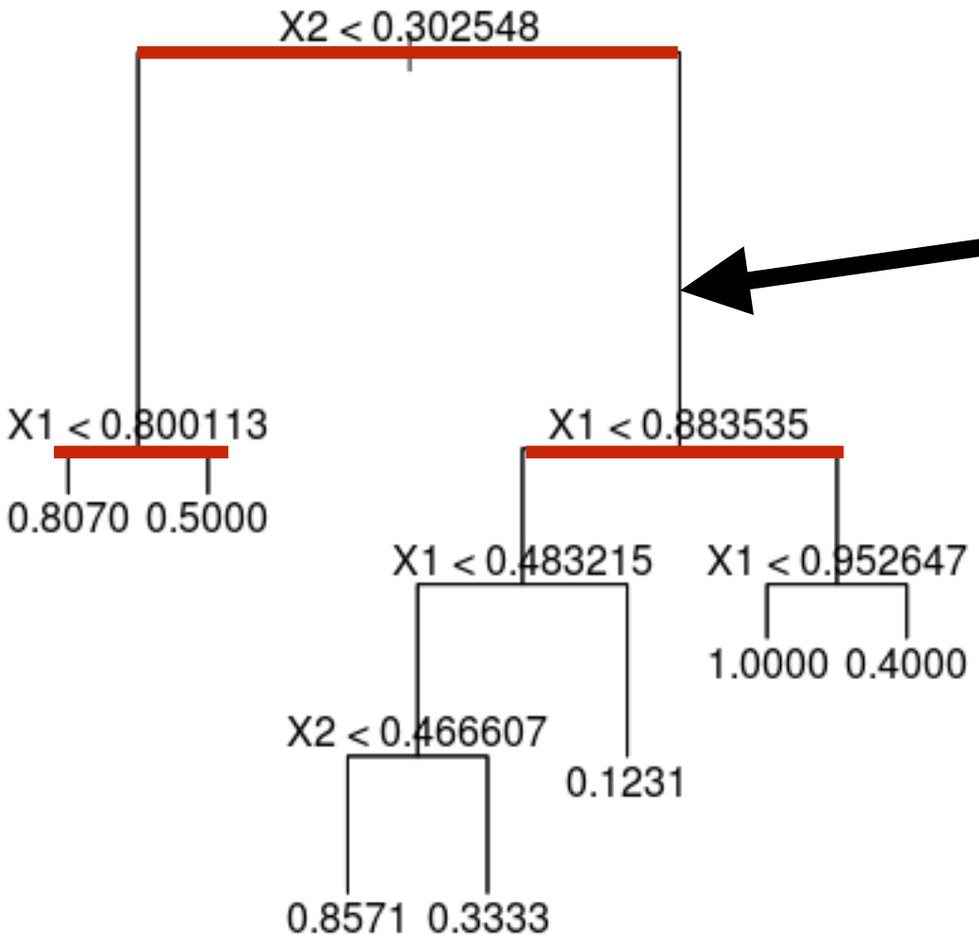


Classification & regression trees





Classification & regression trees

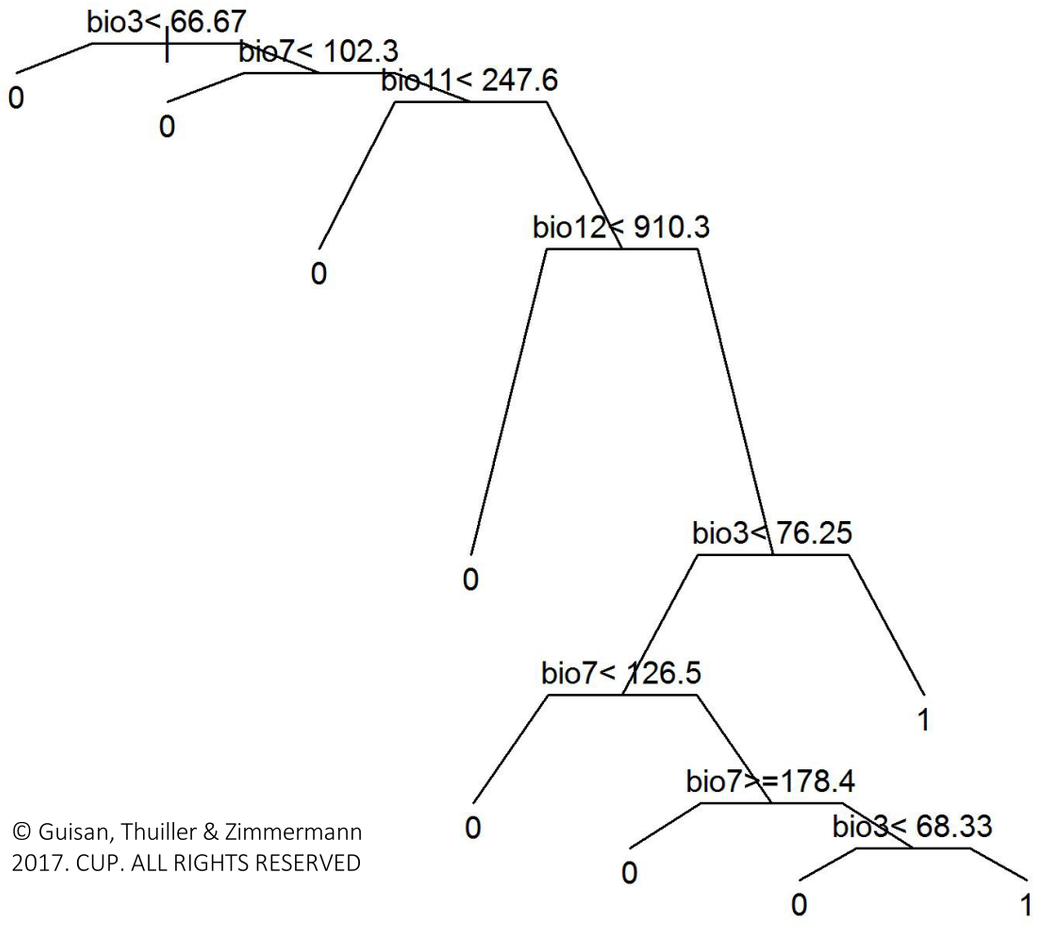


Length of branch is proportional to deviance explained by the split

Classification & regression trees

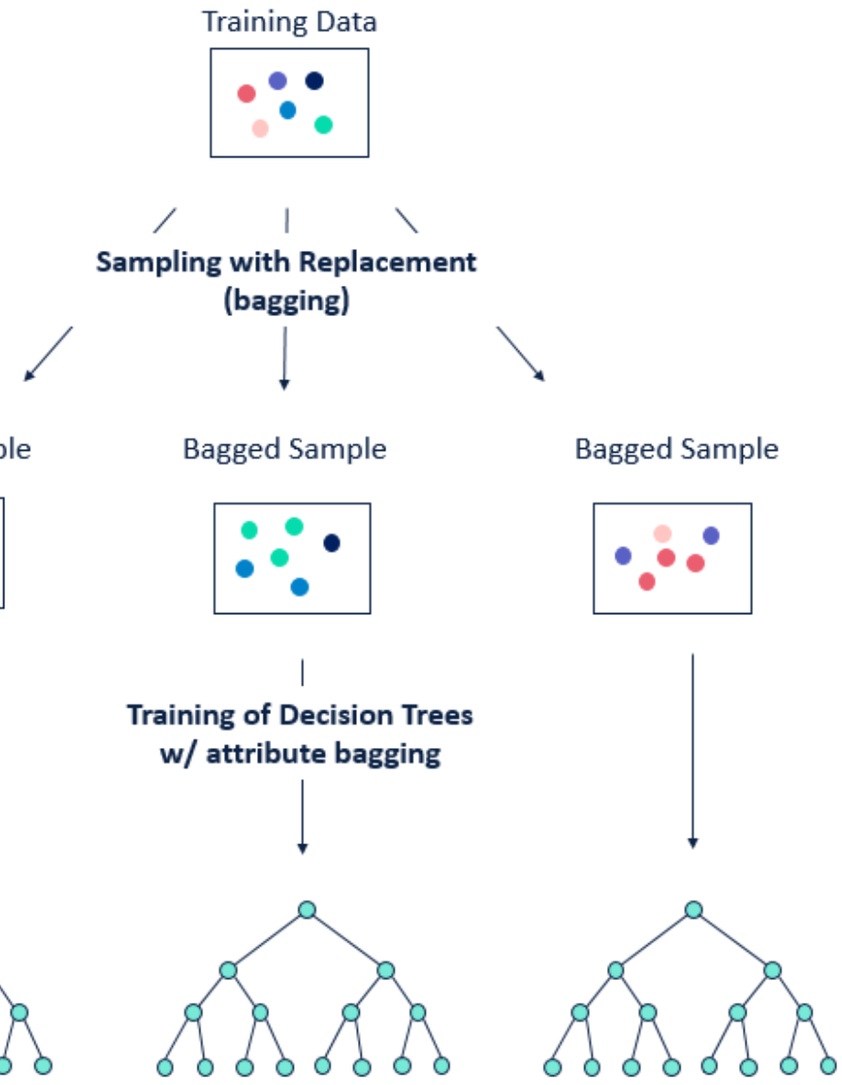
must control model overfitting!!!! otherwise your model will make terrible predictions and not work

- Advantages
  - Categorical predictors
  - Interactions between predictors
  - Threshold responses
  - Missing data
  - Classify new data
  - Informative output



ML: Random Forests (RF)

- Exploit strengths of decisions trees while dealing with weaknesses
- “Bagging” = bootstrap aggregation
- Build a large number of trees and aggregate

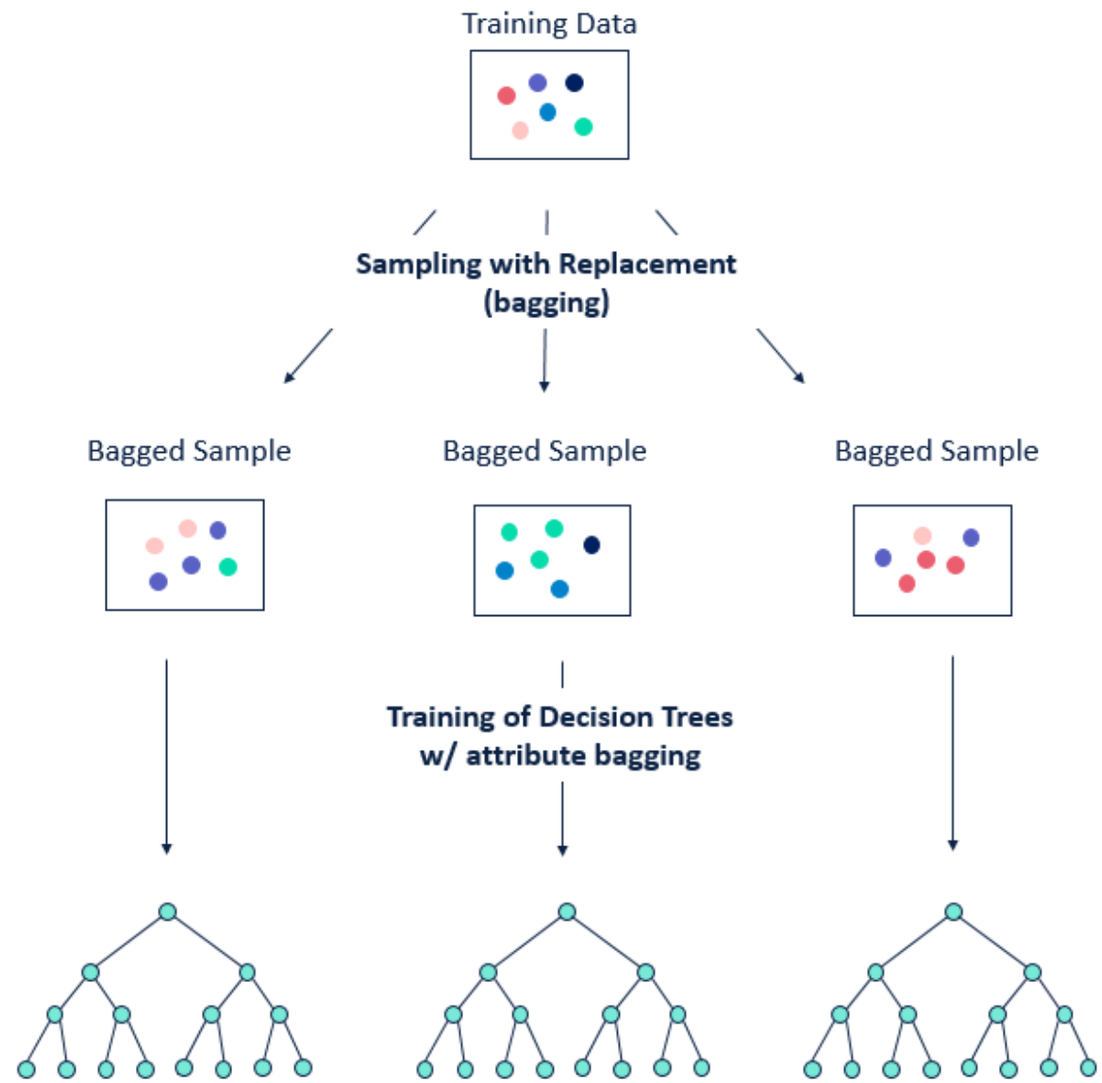


ML: Random Forests (RF)

growing a forest of decision trees and aggregating them at the end

- Steps

1. Subsample data many times with replacement
2. Retain out-of-bag (OOB) sample (~30%) for model evaluation
3. Build a large number of trees (~500), selecting variables at random for each split
4. Evaluate with OOB
5. Avoid overfitting by averaging forest of trees



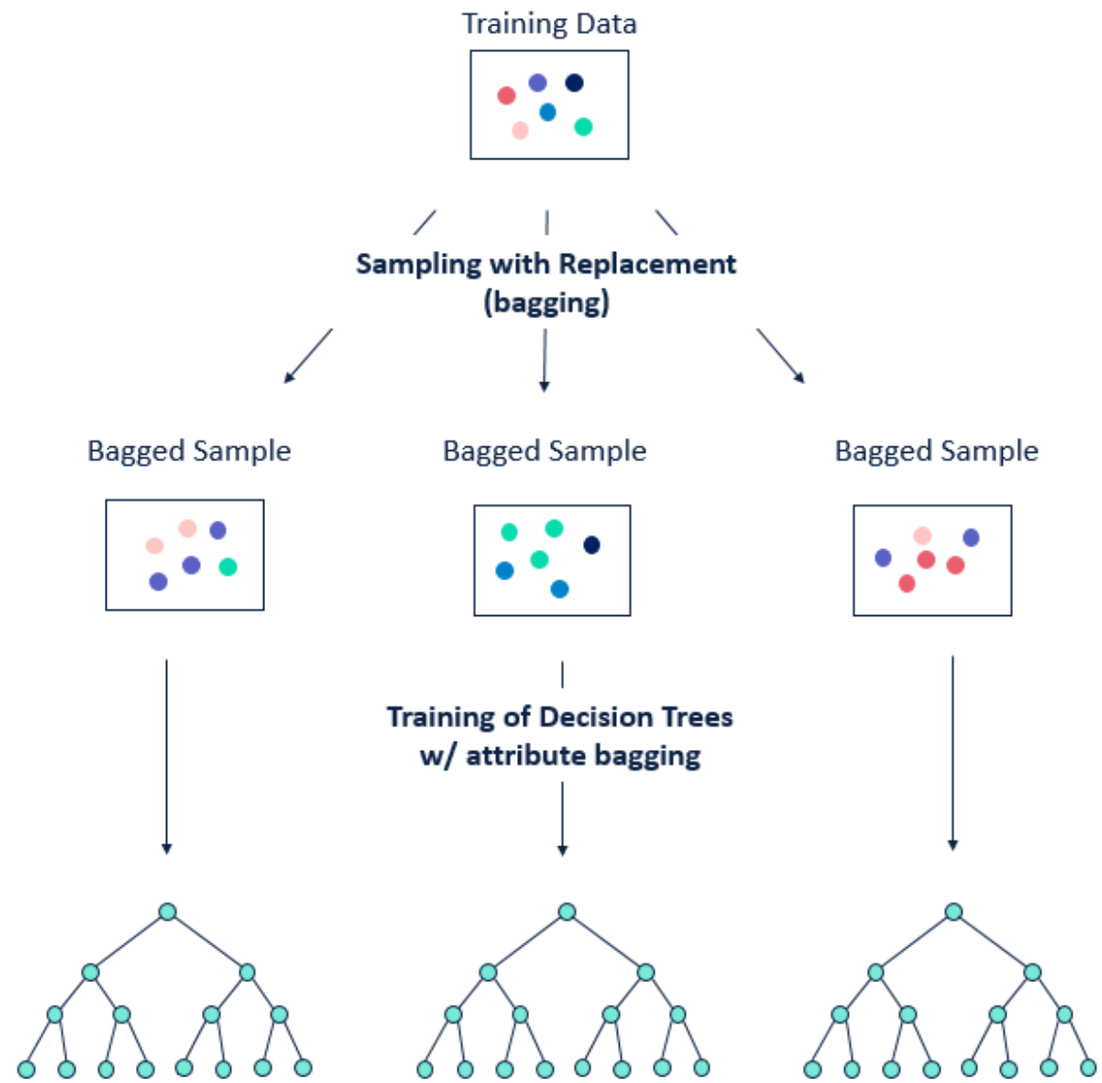
ML: Random Forests (RF)

- Variable importance:

    1. Calculate error using OOB

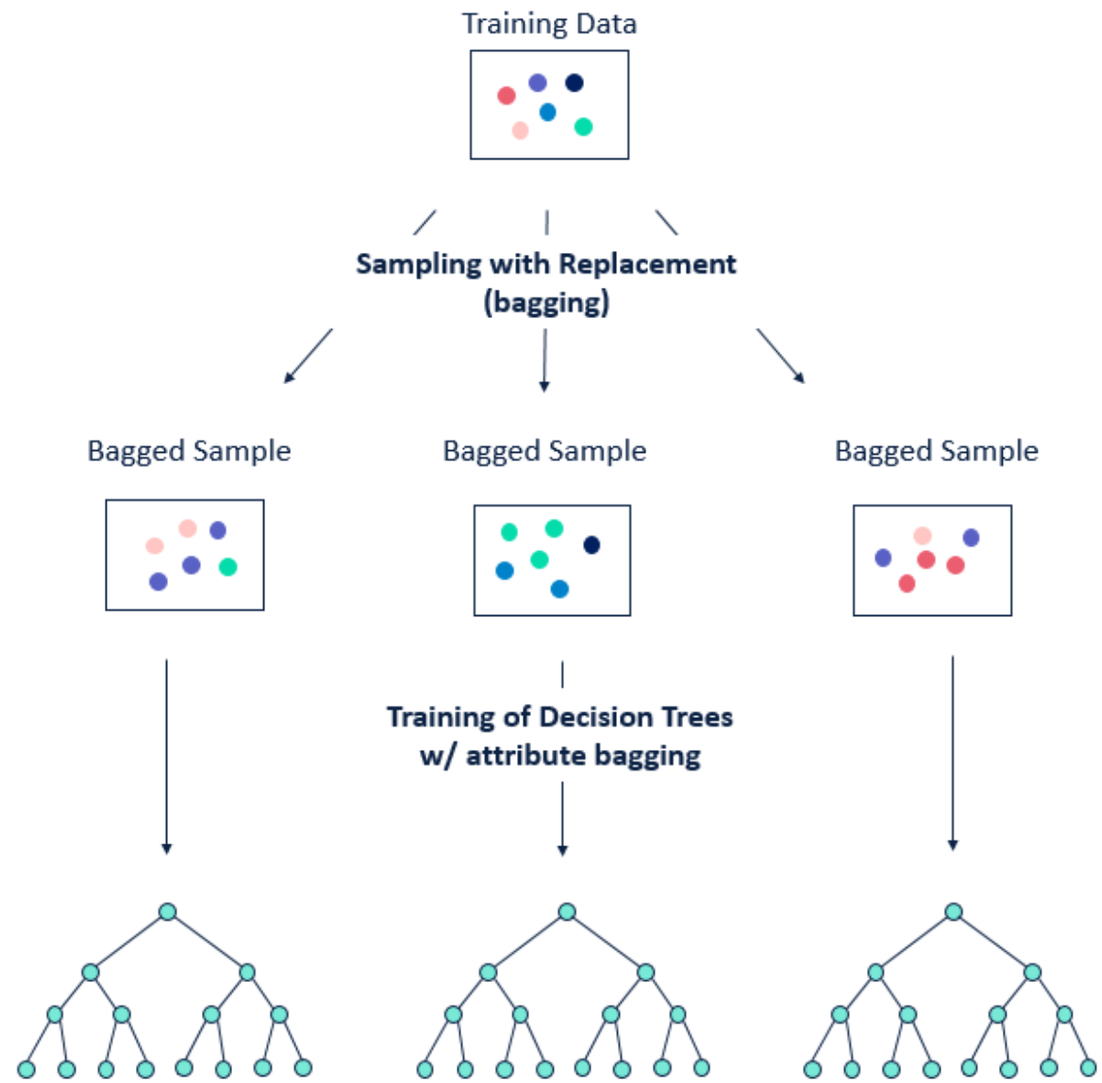
    2. Randomize each variable in turn and calculate how much error rates change

        - Important variables = large increase in error when randomized



ML: Random Forests (RF)

- Variable importance:
  - Calculate the reduction in deviance achieved by all splits in a tree that use that variable, averaged across all trees



ML: Boosted Regression Trees (BRT / GBM)

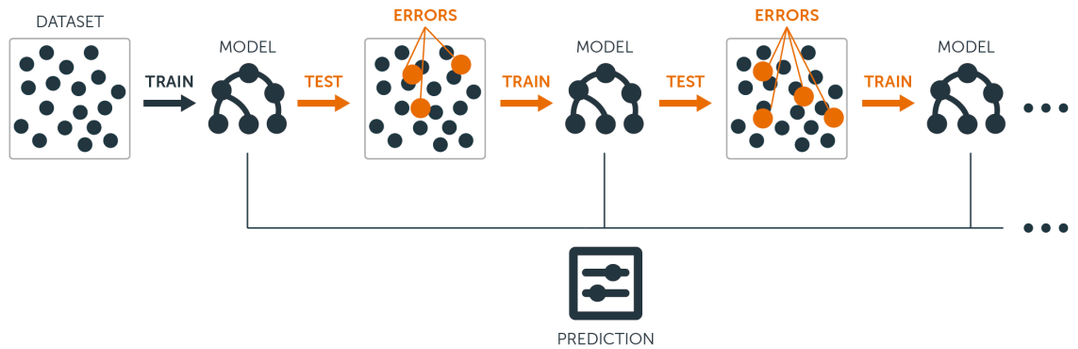
- “Boosting” = adaptive method for combining many simple models to improve predictive performance

- Easier to find and combine many rough rules-of-thumb than to find a single, highly accurate prediction rule

- Combines strengths of statistical and machine learning methods

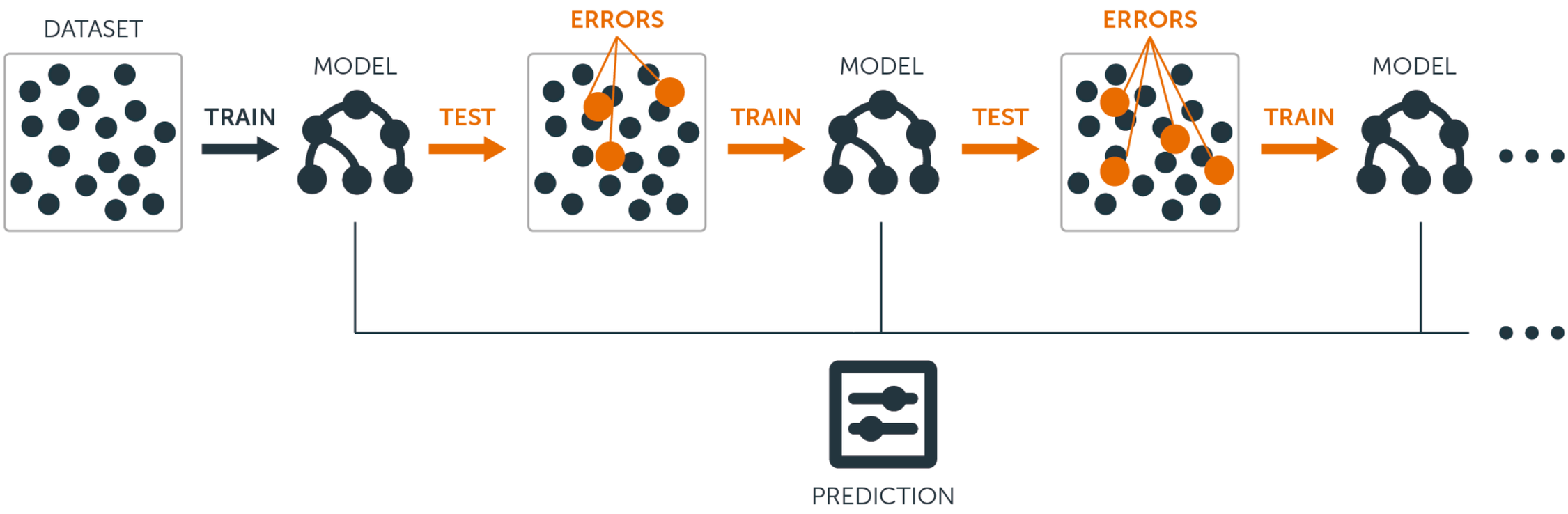
good for many interactions and non-linear relationships

it tends to not suffer from overfitting as much as random forest can



ML: Boosted Regression Trees (BRT / GBM)

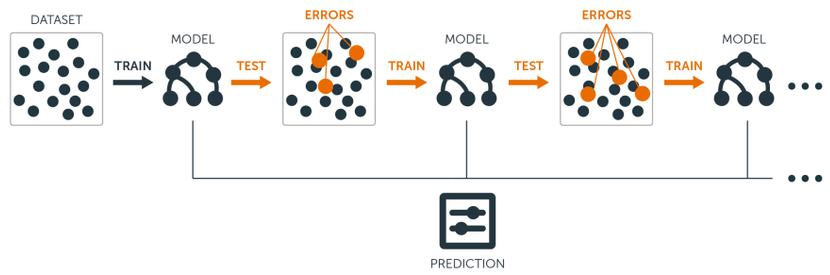
- Sequential algorithm: forward, stage-wise
- As algorithm learns, emphasis is placed on observations that are modeled poorly using the existing set of trees



ML: Boosted Regression Trees (BRT / GBM)

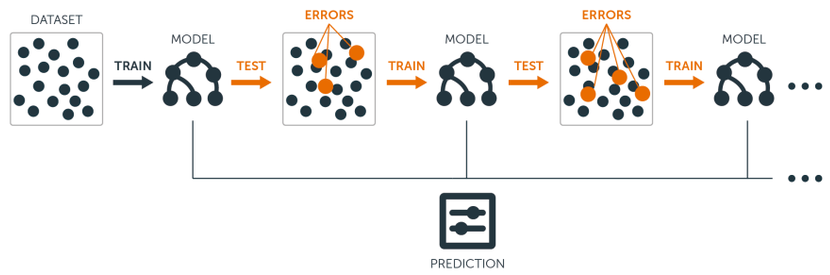
need slower learning rate if your model is very complex and has lots of data

- Final model can be understood as as additive regression model in which coefficients are simple trees
- Contains 100’s to 1000’s of trees



ML: Boosted Regression Trees (BRT / GBM)

- Advantages
  - Very flexible, robust approach with good predictive performance
  - Categorical predictors, interactions, free of certain statistical assumptions
  - Complex, non-linear
  - Robust assessment of variable importance



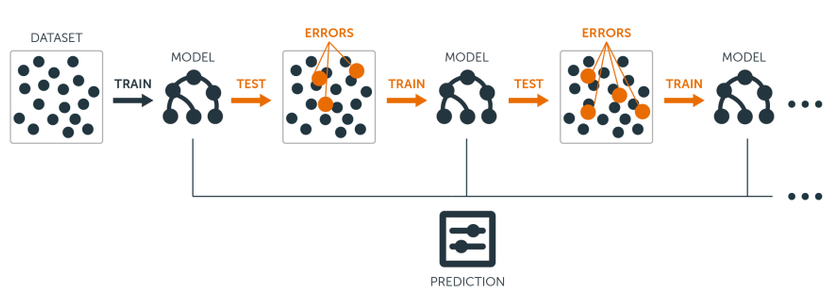
ML: Boosted Regression Trees (BRT / GBM)

GRB= gradient boosting machine
EGB= extreme gradient boosting

- Disadvantages
  - More challenging to fit
    - Computation demands

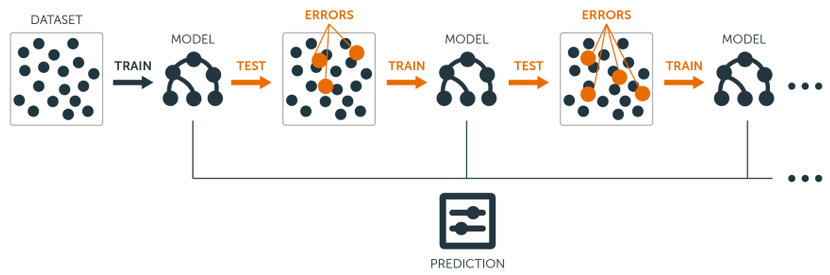
need slower learning rate if your model is very complex and has lots of data

- Tuning of parameters
  - Learning rate (lr)
  - Tree complexity (tc)
- lr & tc ~ number of trees (nt) required for optimal prediction



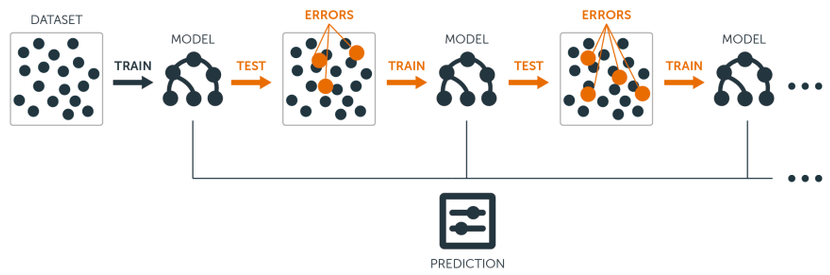
ML: Boosted Regression Trees (BRT / GBM)

- Learning rate (lr)
  - Also known as “shrinkage” parameter
- Determines how much each tree contributes to the growing model
- Lower values = slower learning / smaller contribution
  - Could be better for more complex problems



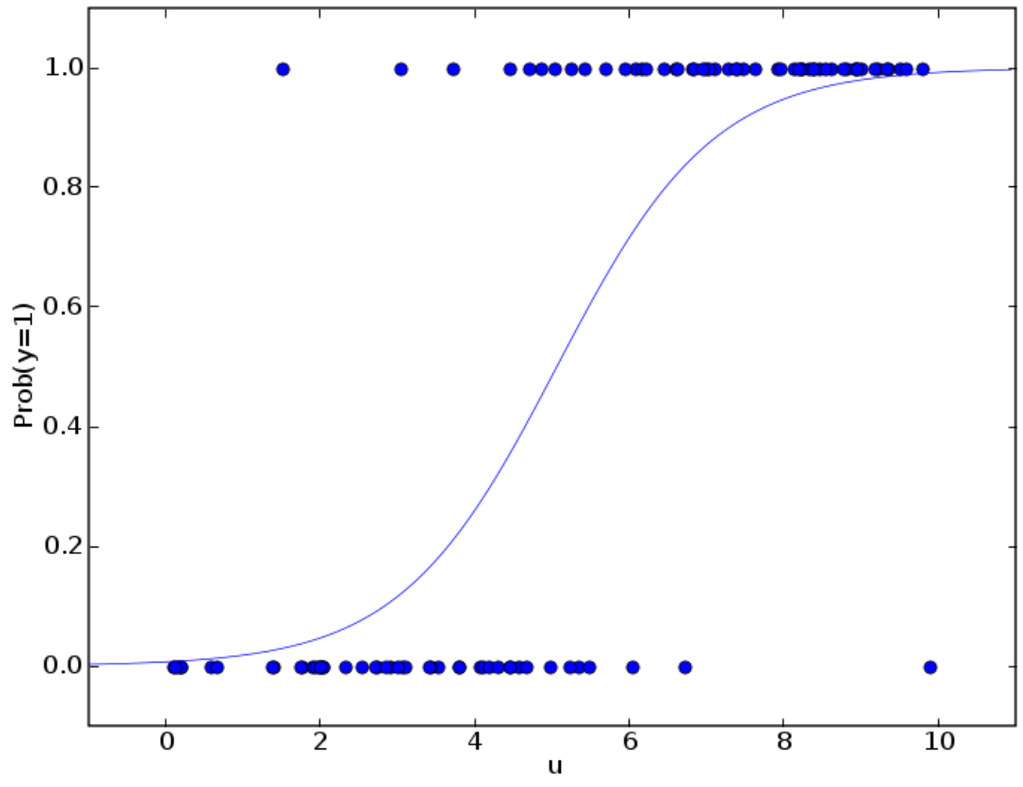
ML: Boosted Regression Trees (BRT / GBM)

- Tree complexity (tc)
  - Controls how many splits each tree can have
  - Determines whether interactions are fitted
  - tc = 1 = additive model
  - tc = 2 = up to two-way interactions



GLM: Generalized Linear Model

- Classic technique widely used in SDM
- Extension of linear regression that can handle non-normal distributions of the response variable
  - Poisson (counts)
  - binomial (presence-absence)

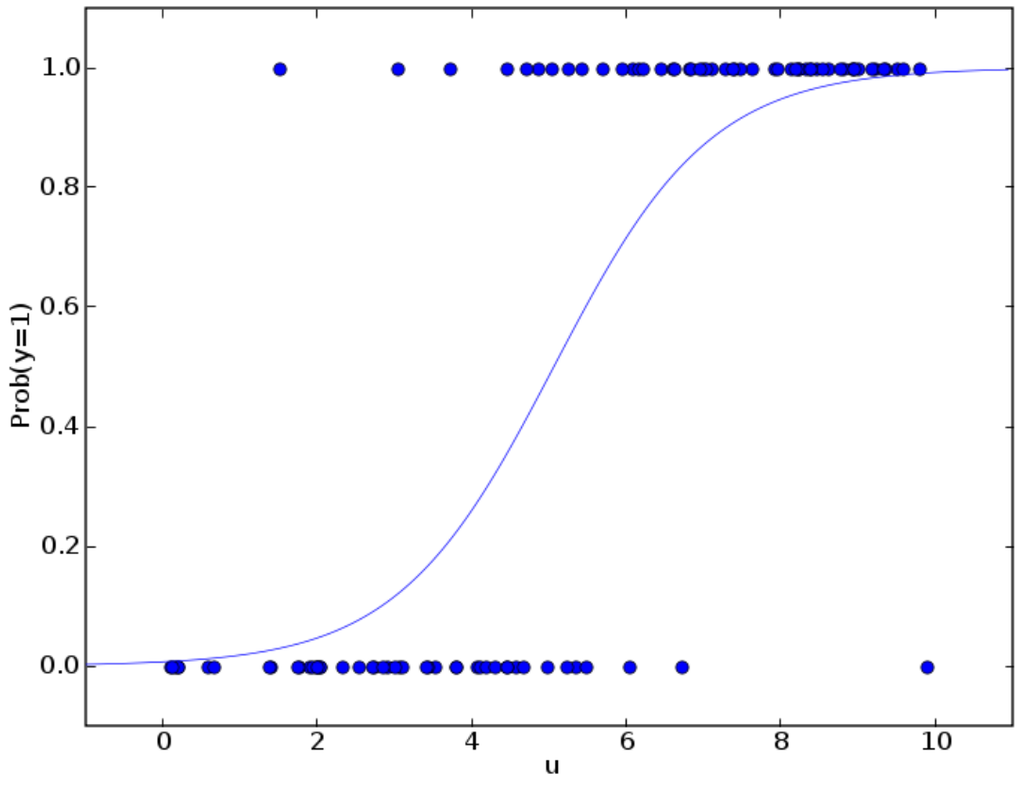


$$log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$$

GLM: Generalized Linear Model

- Must define:

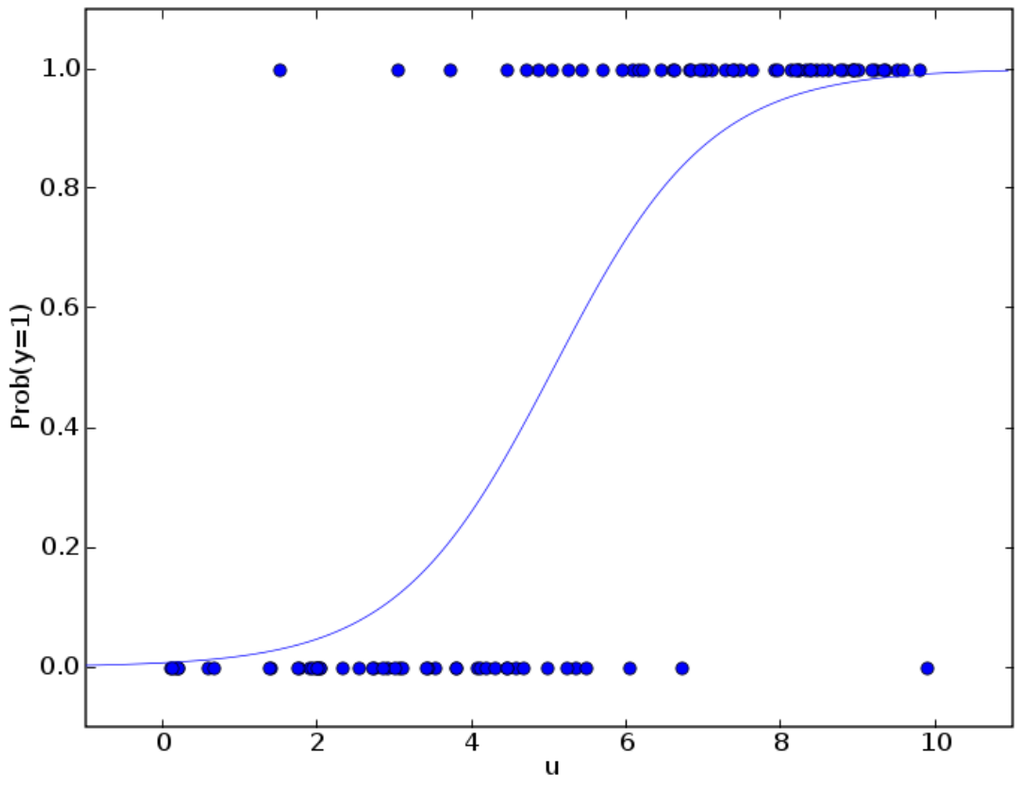
1. Response distribution (e.g., binomial for P-A)
2. Link function
   - Logit-link for P-A
3. Variance function



$$log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$$

GLM: Generalized Linear Model

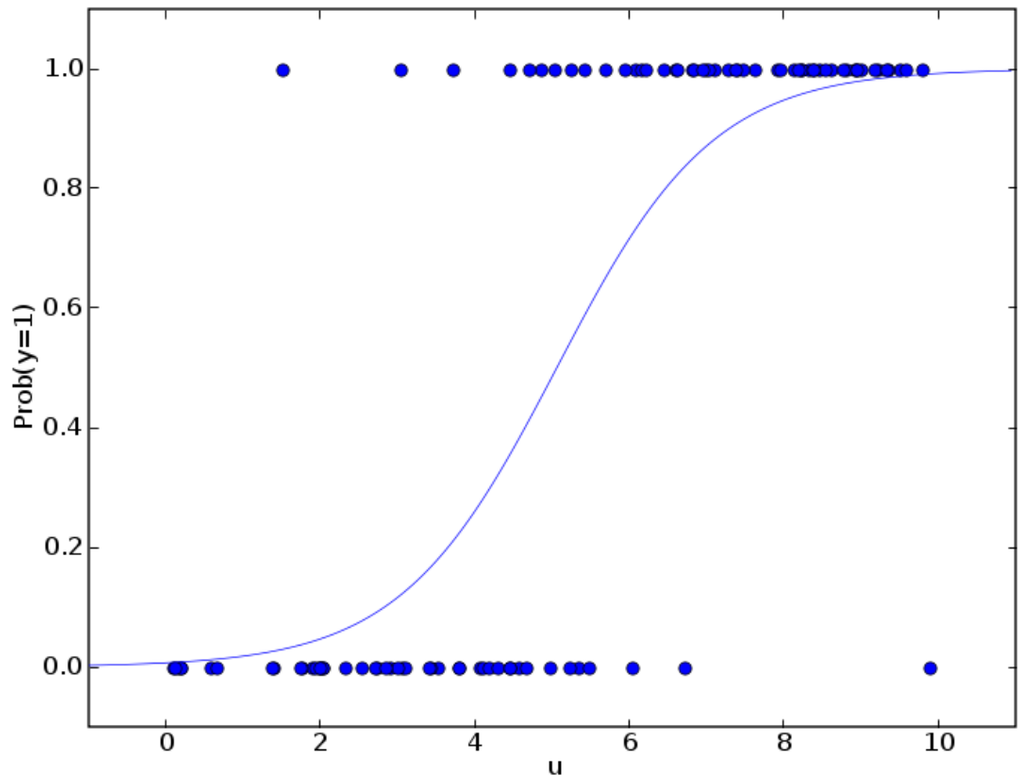
- Coefficients have a convenient interpretation
  - If beta = 0.0249 for temperature
  - a 1 unit increase in temperature results in:
  - exp(0.0249) = 1.025 = 2.5% increase in probability of presence



$$log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j\hat{\beta}_j + \varepsilon$$

GLM: Generalized Linear Model

- Predictions are on the scale of the link function
- For logit = log odds
- **Apply inverse transform to predict in unit of the response variable**
- Non-linear responses can be modeled using polynomials
- AIC can be used for model selection



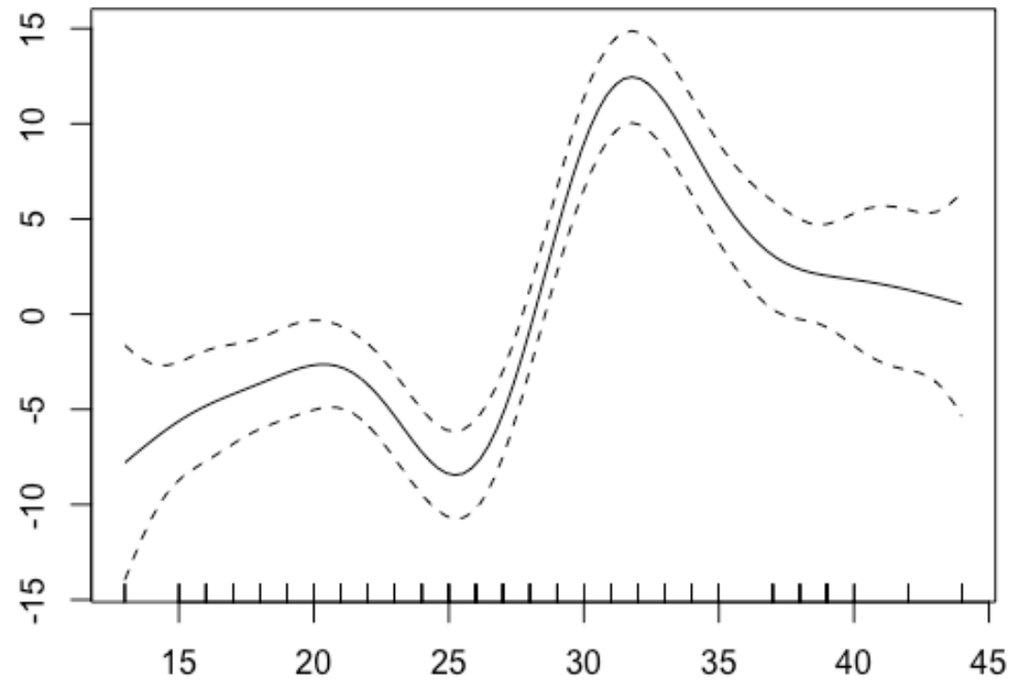
$$log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j + \varepsilon$$

GAM: Generalized Additive Model

- Highly flexible, nonlinear, nonparametric extension of GLM

- Coefficients of GLM are replaced with a smoothing function

- Good for characterizing nonlinear response curves



$$log\left(\frac{\mu}{1-\mu}\right) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j f_j + \varepsilon$$

Questions?