Introduction to algorithms for species distribution modelling

From: Richard Pearson

Adapted by Alison Cameron

Outline:

- Context: the algorithm as just one part of the model
- Review of some common techniques: presence-only and presence-absence approaches
- Model-based uncertainty…

The main steps to build & validate a species distribution model (SDM)

L2

Observed species' distribution (a list of localities where the species has been observed, and sometimes also localities where the species is known to be absent)

L3 & P1, P2

Modeling algorithm e.g. Maxent, GARP, BioClim, Domain, artificial neural network, generalized linear model, regression trees

Predicted species' distribution.
Prediction may be for a different region (e.g. for an invasive species) or for a different time period (e.g. under future climate change)

Case Studies

Database of 'raw' environmental variables (e.g. temperature, precipitation, soil type).

Data usually stored in a GIS

L2

Processing to generate variables of importance in defining species' distributions (e.g. maximum daily temperature, frost days, soil water balance)

Model testing (statistical assessment of predictive ability, using test such as AUC or Kappa)

L4, L5, P1, P2

Environmental Niche Modeling (ENM)

Often called **Species Distribution Modeling (SDM)**

The use of **computer algorithms** to:

- **FIRST** fit **mathematical functions,** describing species distributions in **environmental space,** for each of n environmental variables
- **THEN** generate **predictive maps** of species distributions in **geographic space** using these functions

Remember: the algorithm as just one part of the modeling process…

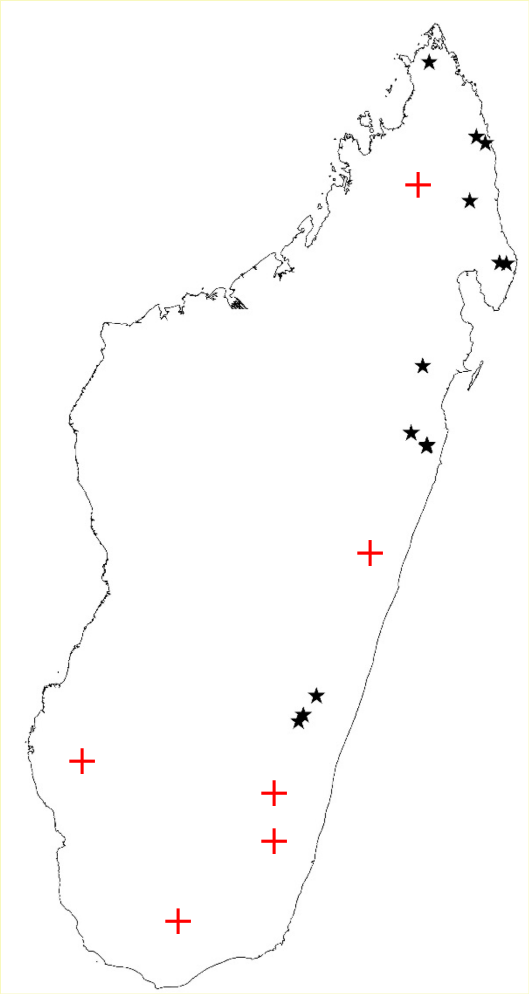
1. Definition of the question
2. Data identification and preparation (environmental layers, presence only vs presence/absence, scale, data partitioning)
3. Selection of a modeling algorithm (absence data, categorical data, extrapolation)
4. Testing predictive performance (omission/commission errors, decision thresholds, variable contributions)
5. Interpretation of model output

Some approaches that have been applied:

Method(s)	**Model/software name**	**Species data type**
Climatic envelope	BIOCLIM	Presence-only
Gower Metric	DOMAIN	Presence-only
Ecological Niche Factor Analysis (ENFA)	BIOMAPPER	Presence/background
Maximum Entropy	MAXENT	Presence/background
Genetic algorithm	GARP	Presence/pseudo-absence
Regression: Generalized linear model (GLM) and Generalized additive model (GAM)	GRASP	Presence/absence
Artificial Neural Network (ANN)	SPECIES	Presence/absence
Classification and regression trees (CART), GLM, GAM and ANN	BIOMOD	Presence/absence
Boosted regression trees	BRT (*implemented in R)*	Presence/absence
Multivariate adaptive regression splines	MARS (*implemented in R)*	Presence/absence
Ensembles	*SDMtoolbox (Brown, 2014)* *BIOMOD*	**??????**

Species' distribution data: presence-only or presence/absence?

Uroplatus sp.
(leaf-tailed gecko)



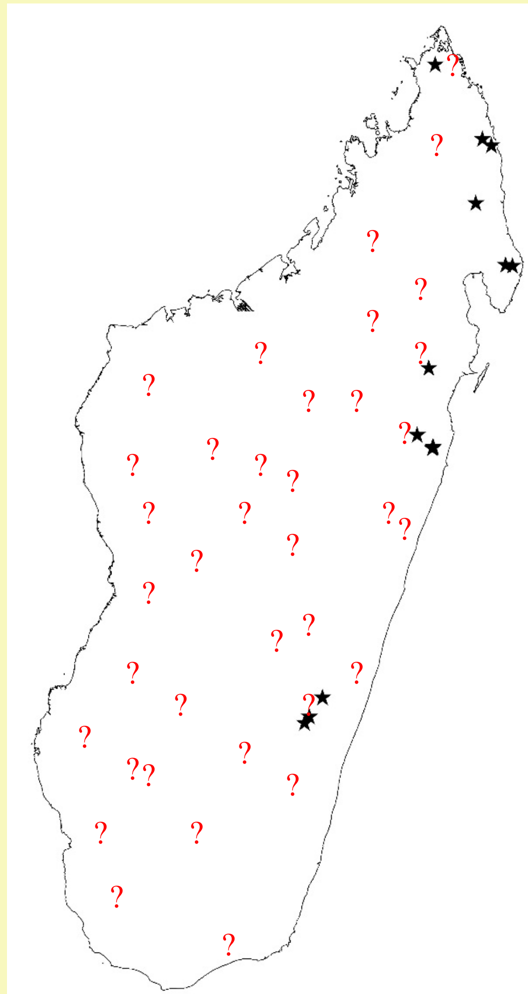
★ Observed 'presence' record

+ Observed 'absence' record*

(*absence records on this map are hypothetical)

Species' distribution data: presence-only or presence/absence?

Uroplatus sp.
(leaf-tailed gecko)

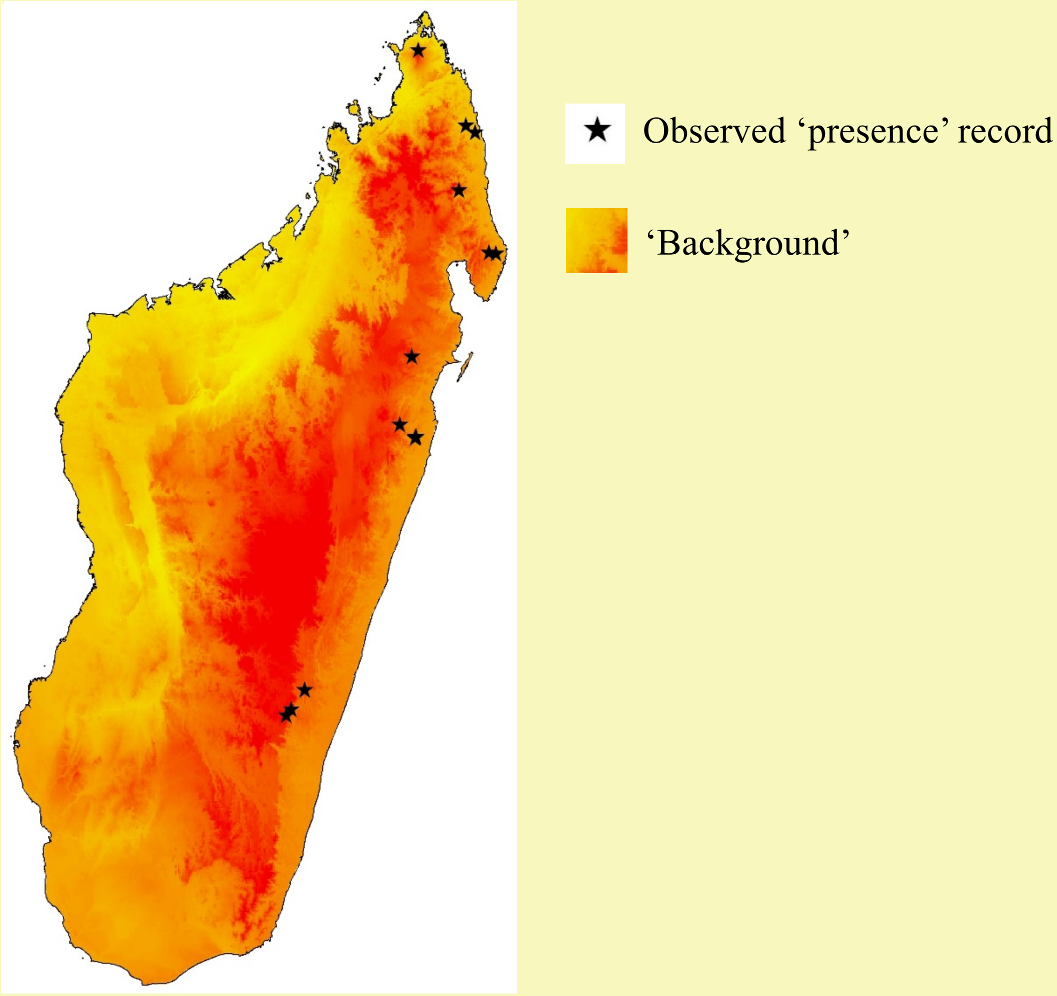


★ Observed 'presence' record

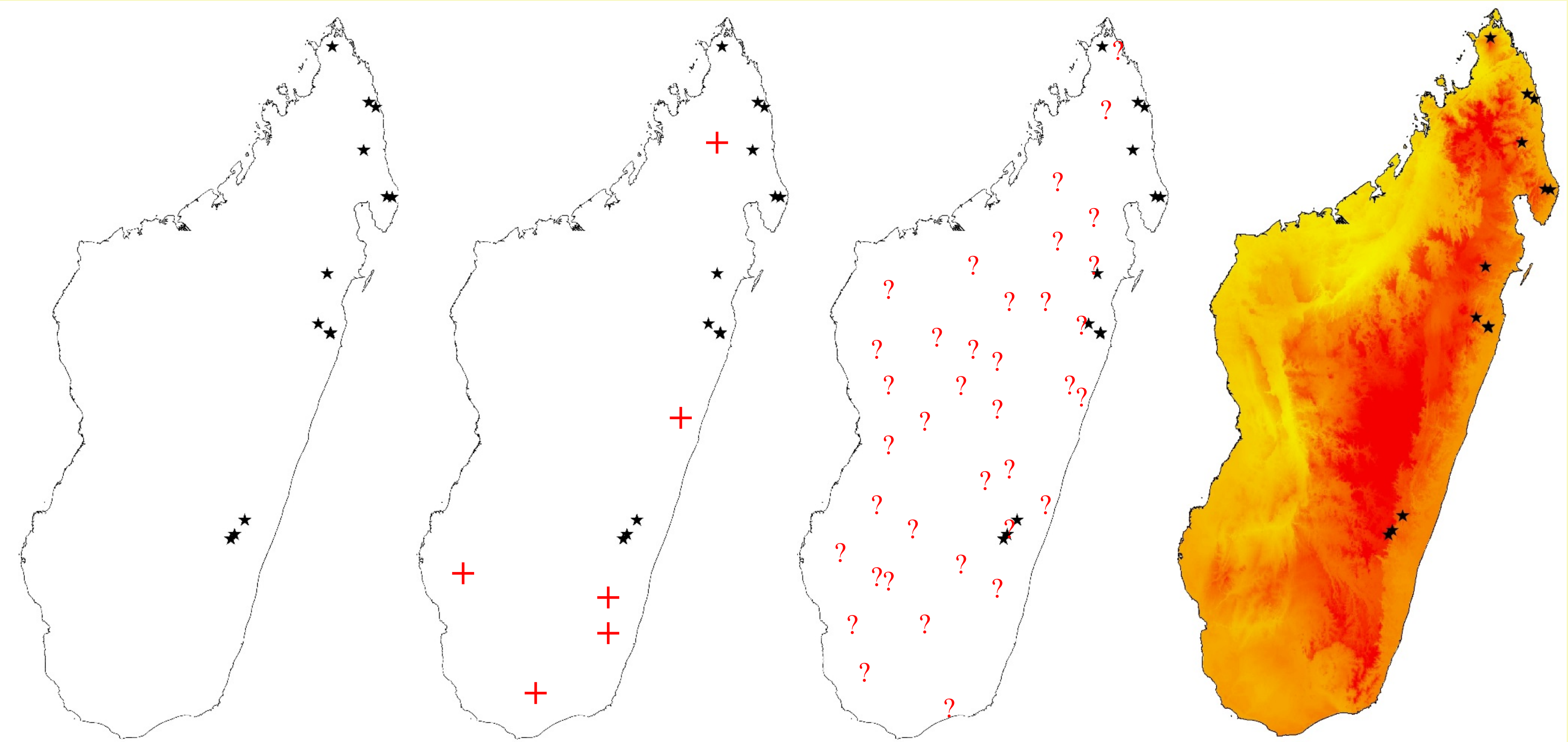
? 'Pseudo-absence'

Species' distribution data:
presence-only or presence/absence?

Uroplatus sp.
(leaf-tailed gecko)



Species' distribution data:



Presence-only

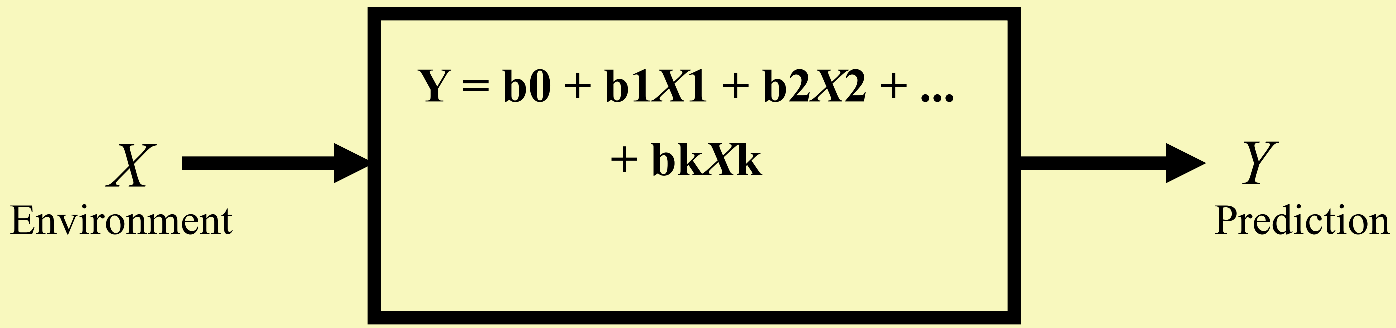
Presence/absence

Presence/pseudo-absence

Presence/background

General consideration: Explanation or prediction?

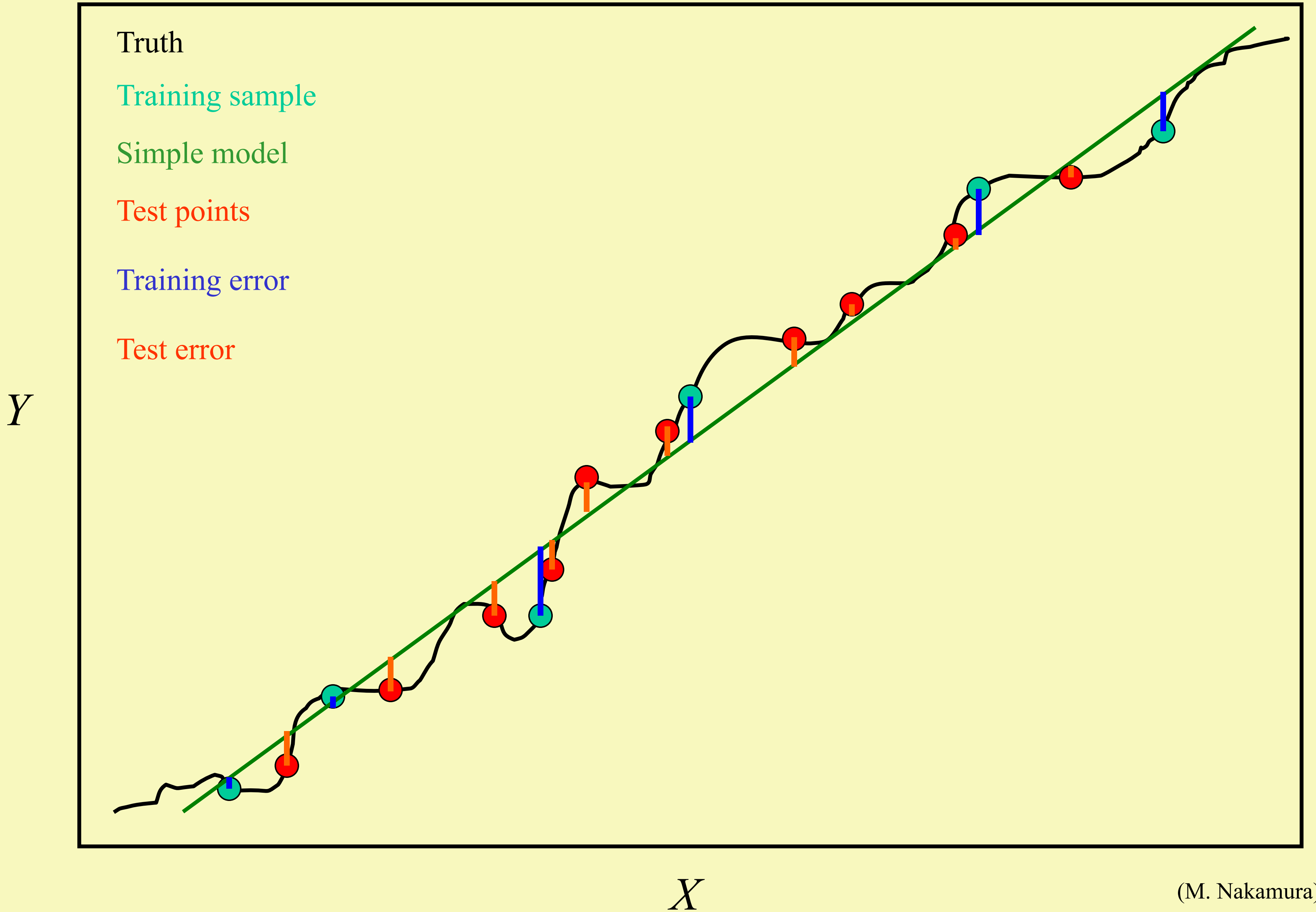
- Explanation/understanding:
  - a simple approximation often preferred (Occam's razor). E.g. GLM.
  - More complex models: explain with visualization, response curves
- Prediction:
  - Best possible approximation usually preferred. May be complex, difficult to interpret -- may not help us understand the system.



or black box?

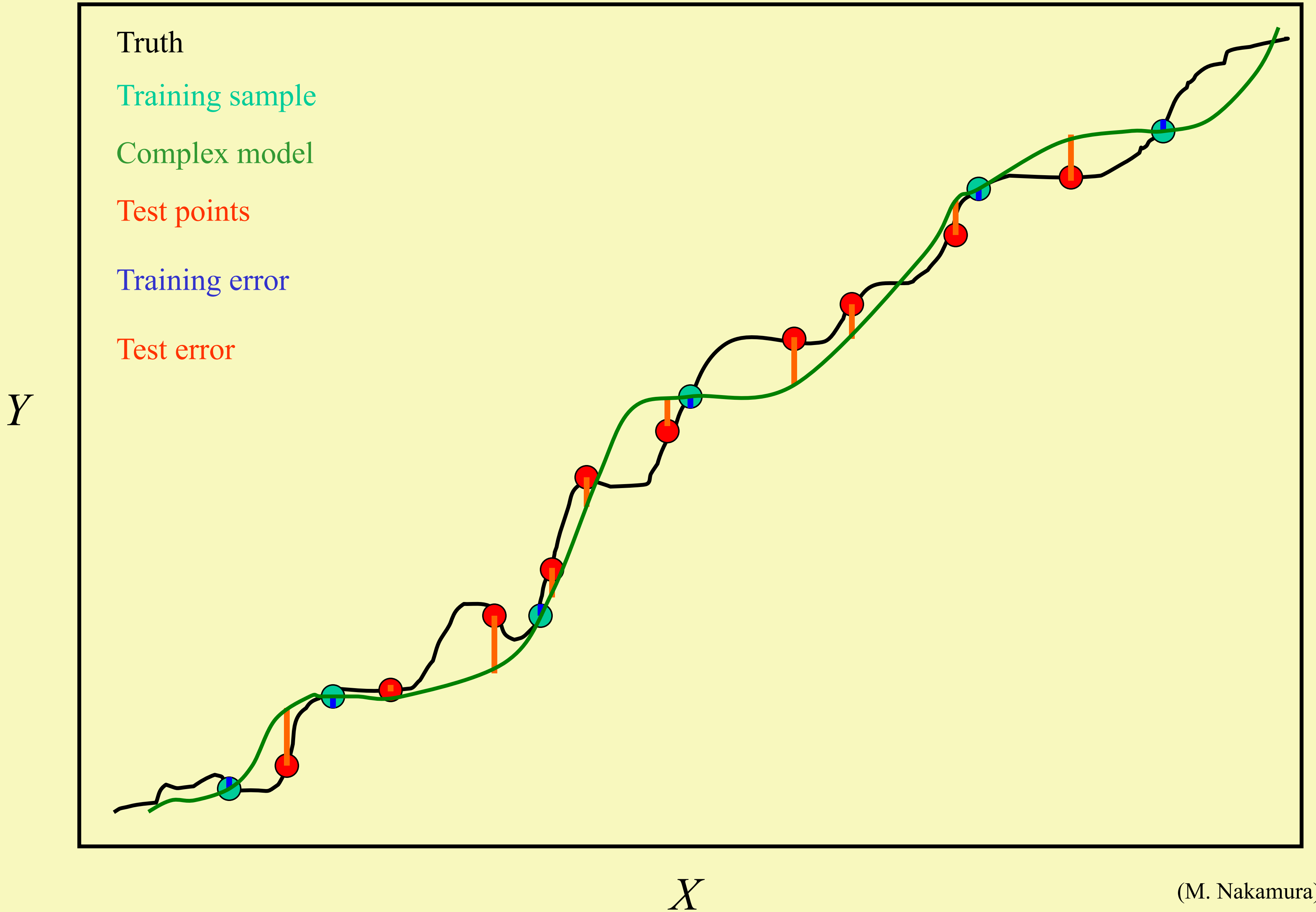
(M. Nakamura)

General consideration: model complexity



(M. Nakamura)

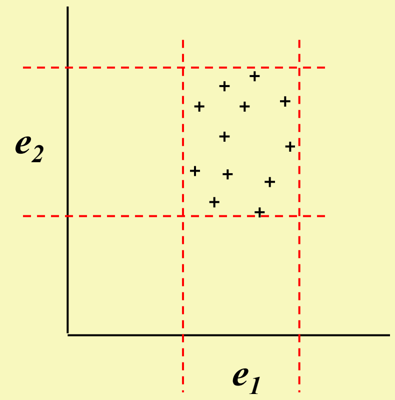
General consideration: model complexity



(M. Nakamura)

Algorithm using only presence records

Climate Similarity: BIOCLIM



- Simple and intuitive similarity model
- Gives equal weight to all variables
- Does not account for potential interactions between variables
- Gives binary predictions (or continuous, defined by minimum percentile)
- Cannot use categorical variables
- No extrapolations into “novel” conditions

See: Nix 1986… or Lindenmayer et al. 1991 *J. Biogeog.* 18: 371-383.
Arcscript: http://arcscripts.esri.com/details.asp?dbid=13745
Diva GIS: http://diva-gis.org

Algorithms using presence and absence records

Regression: Generalized linear model (GLM) and

Generalized additive model (GAM)

• Implemented in SPLUS and R by the 'Generalized Regression Analysis and Spatial Prediction' group (GRASP)

• 'Transparent' statistical approaches

• GLMs assume a linear relationship between (transformed) response and predictors

See Guisan et al. 2002 *Ecological Modeling* 157: 89-100

Lehman et al. 2002 *Ecological Modeling* 157: 189-207

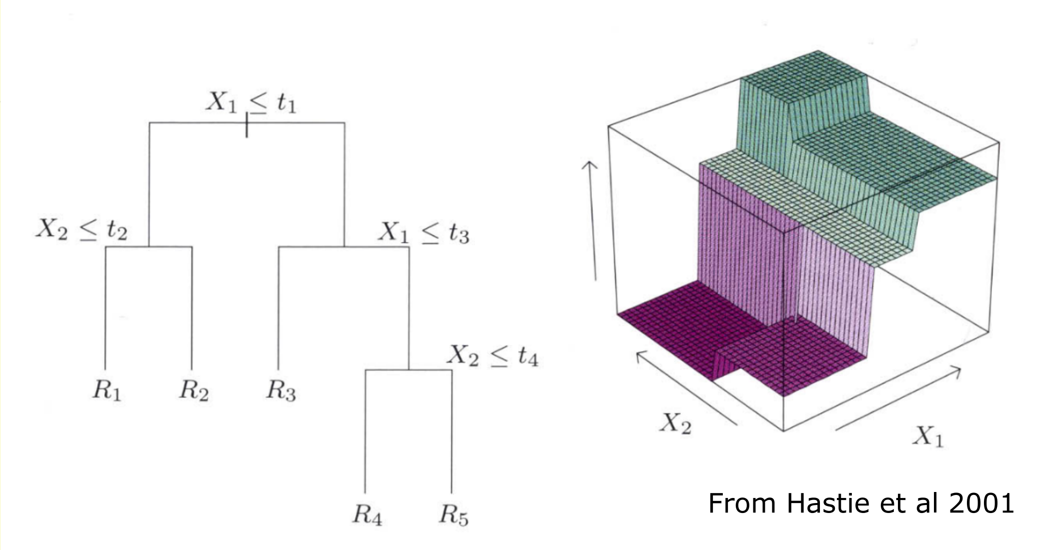
Algorithms using presence and absence records

Regression trees

- Classification and regression trees, **CART**
- Recursive binary splits
- Grow large and prune / fixed size

✓Good:

- Numeric vars, categorical, ..
- Variable selection
- Outliers
- Transformations
- Missing data
- Interactions



From Hastie et al 2001

✗Bad:

- Inaccurate

Algorithms using presence and absence records:

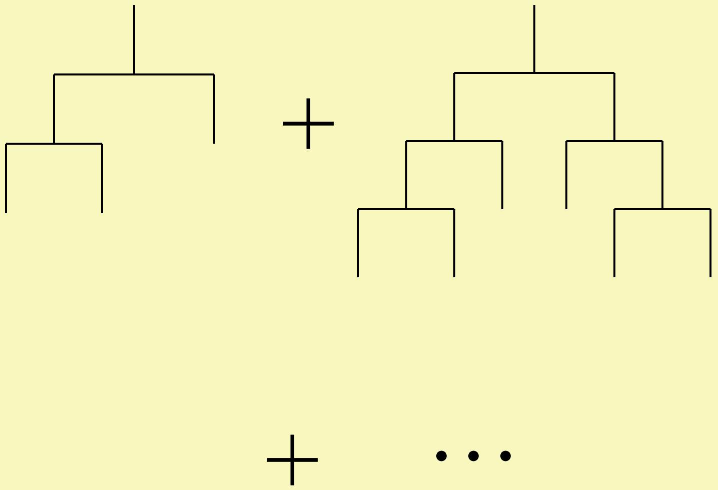
Boosted or bagged regression trees

✓Random Forests:

- Sequence of trees fitted independently
- Output is sum of trees
- Salford Systems

✓BRT:

- Boosted regression tree / stochastic gradient boosting
- Fit sequence of trees
- Output is sum of trees
- Each tree fitted to improve fit of trees so far
- Free implementation in R



See: Elith et al, 2008 Working Guide to Boosted Regression Trees

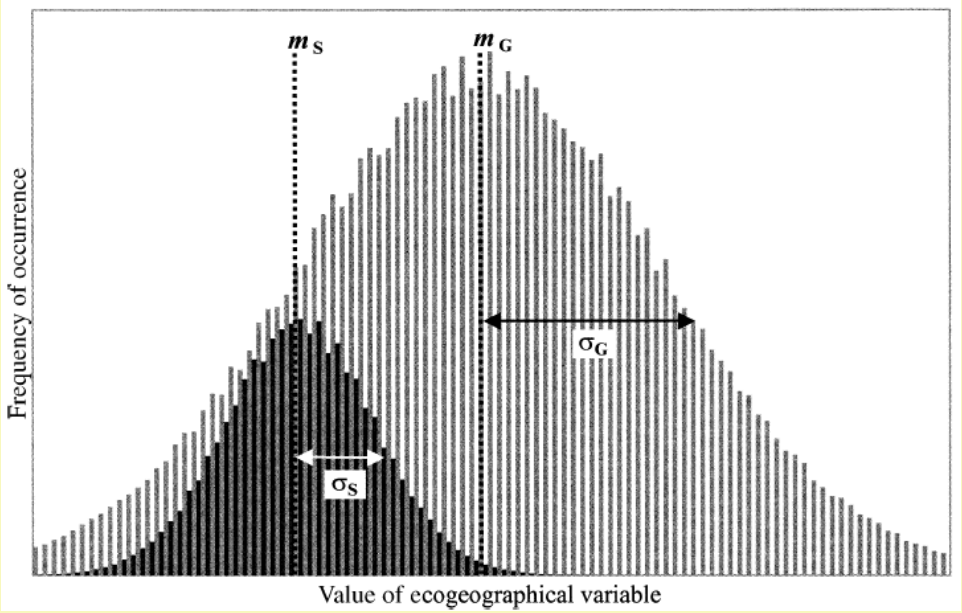
Algorithms using presence & background data

MAXENT: more to come…

ENFA (ecological niche factor analysis)

- Biomapper implementation: http://www2.unil.ch/biomapper/
- Cannot interpret categorical (discrete) input

See: Hirzel et al. 2002 *Ecology* 83: 2027-2036.



$$\text{Marginality} = \frac{|m_G - m_S|}{1.69\sigma_G}$$

$$\text{Specialization} = \frac{\sigma_G}{\sigma_S}$$

Algorithm using presence & pseudo-absences

Genetic algorithm for rule-set prediction: GARP

- Uses a genetic algorithm to develop rules based, in part, on climate envelopes and general linear models

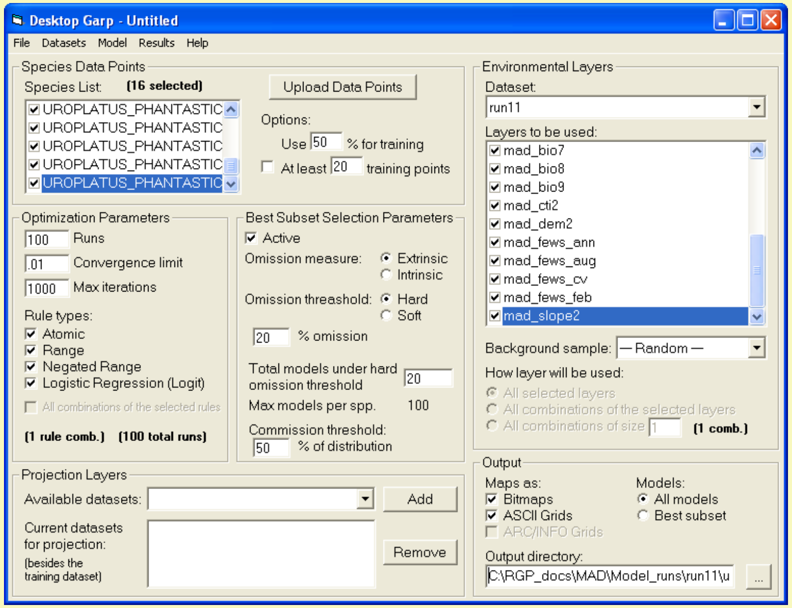
•Output is a sum of runs

•Neat user interface

- Widely applied to address a range of questions

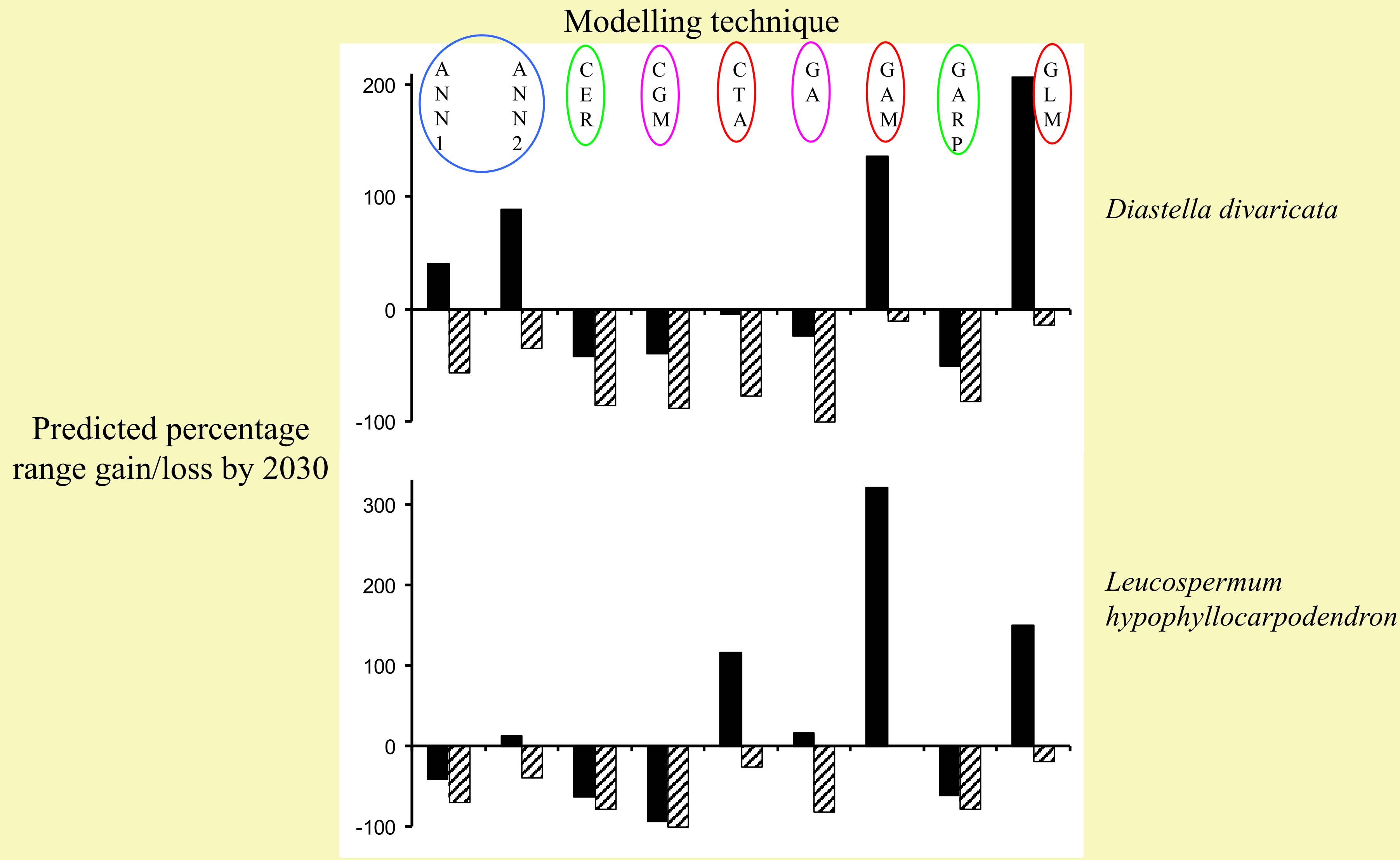
- Computationally intensive

- Poor at interpreting categorical data



See: Stockwell and Peters 1999 *Int. J. Geographical Info. Systems* 13: 143-158; Anderson et al. 2003 *Ecological Modelling* 162: 211-232
… and papers by A. Town Peterson and colleagues
http://www.lifemapper.org/desktopgarp/

How important is model method? Model-based uncertainty



(Pearson *et al., J. Biogeog.* 2006; see also Thuiller *et al*. *Nature*, 2004)

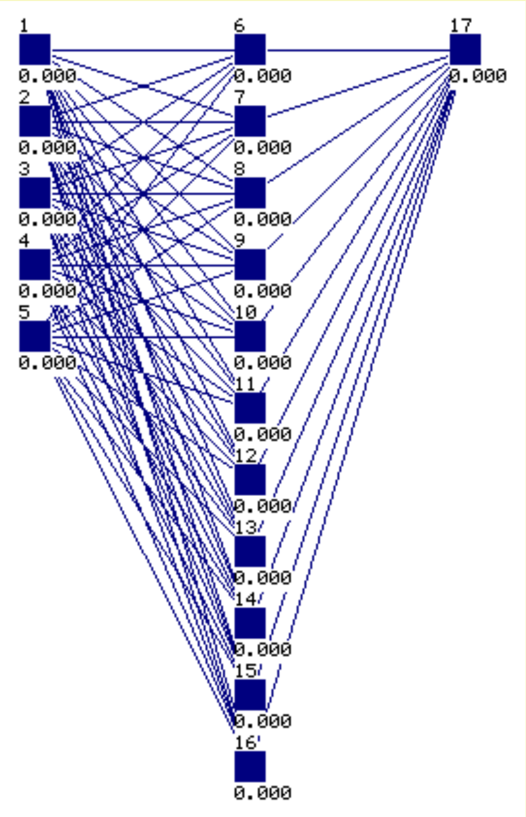
Other algorithms/models (group contributions??)…

- Fuzzy envelope: Svenning & Skov
- Mahalonobis distance: pres-background; no categorical; ArcView extension.
- WhyWhere: David Stockwell
- Domain
- Lives; support vector machines; (co)kriging

Algorithms using presence & absence records

Artificial Neural Network (ANN)

- An machine-learning approach, inspired by the structure of the brain
- Theoretically good at identifying non-linear relationships, and robust to noise
- Network structure is difficult to interpret, making the approach fairly ‘black box’
- Can be adapted to interpret categorical data
- Various software packages available; but recently implemented for distribution modeling in software by Oxford University group (SPECIES model)



See: Pearson et al. 2004 *Ecography* 27: 285-298
Hilbert and Ostendorf 2001 *Ecological Modelling* 146: 311-327