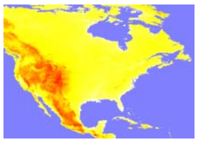
Maximum entropy modeling of species geographic distributions

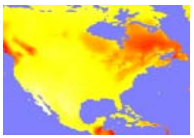
Steven Phillips

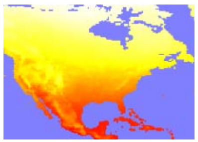
with Miro Dudik & Rob Schapire

Modeling species distributions

occurrence points



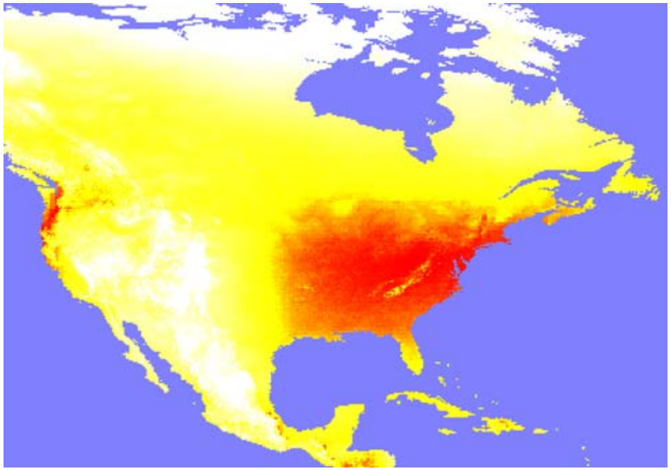




. . .

environmental
variables

Yellow-throated
Vireo



Predicted distribution

Estimating a probability distribution

Given:

- Map divided into cells
- Environmental variables, with values in each cell
- Occurrence points: samples from an unknown distribution

Our task is to estimate the unknown probability distribution

Note:

- The distribution sums to 1 over the whole map
- Different from estimating probability of presence
- $\Pr(t|y{=}1)$ instead of $\Pr(y{=}1|\boldsymbol{x})$ (t=cell, y=response, $\boldsymbol{x}$=environ)

The Maximum Entropy Method

Origins: Jaynes 1957, statistical mechanics

Recent use:

machine learning, eg. automatic language translation

macroecology: SAD, SAR (Harte et al. 2009)

To estimate an unknown distribution:

1. Determine what you know (constraints)
2. Among distributions satisfying constraints:

   Output the one with maximum entropy

Entropy

More entropy : more spread out, closer to uniform distribution

2nd law of thermodynamics:

- Without external influences, a system moves to increase entropy

Maximum entropy method:

- Apply constraints to remove external influences
- Species spreads out to fill areas with suitable conditions

Using Maxent for Species Distributions

"Features"

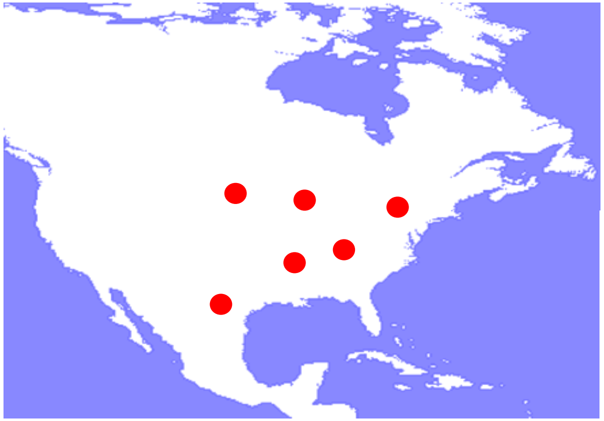
"Constraints"

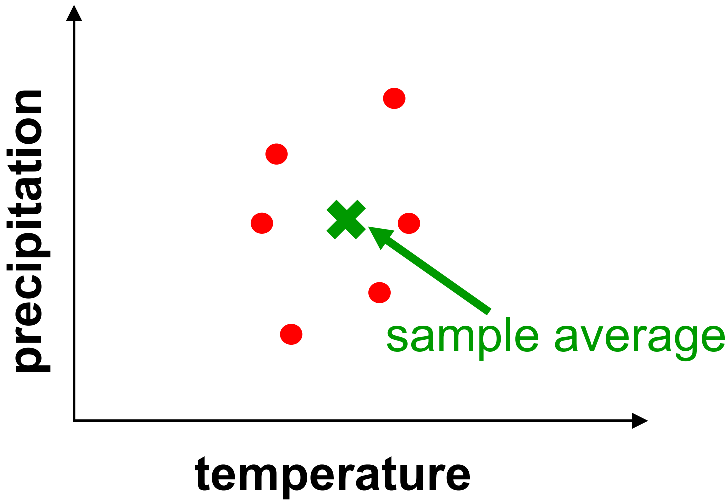
"Regularization"

Free software:
www.cs.princeton.edu/~schapire/maxent/

Features impose constraints

Feature = environmental variable, or function thereof





find distribution of maximum entropy such that
for all features **f**: mean(**f**) = sample average of **f**

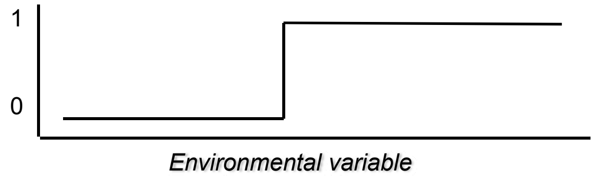
Features

Environmental variables or simple functions thereof.

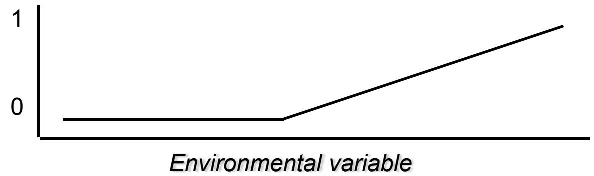
Maxent software has these classes of features (others are possible):

1. Linear … variable itself
2. Quadratic … square of variable
3. Product … product of two variables
4. Binary (indicator) … membership in a category

5. Threshold …



6. Hinge …

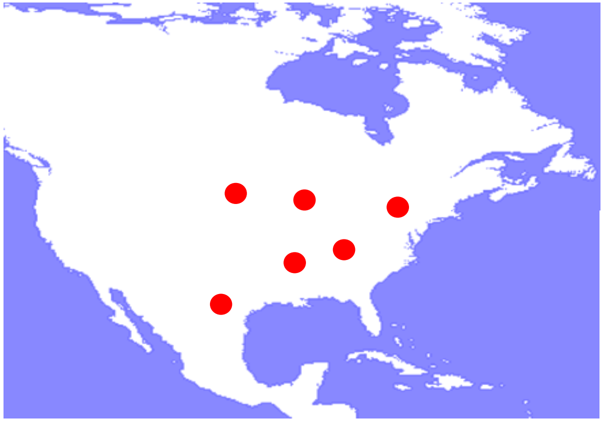


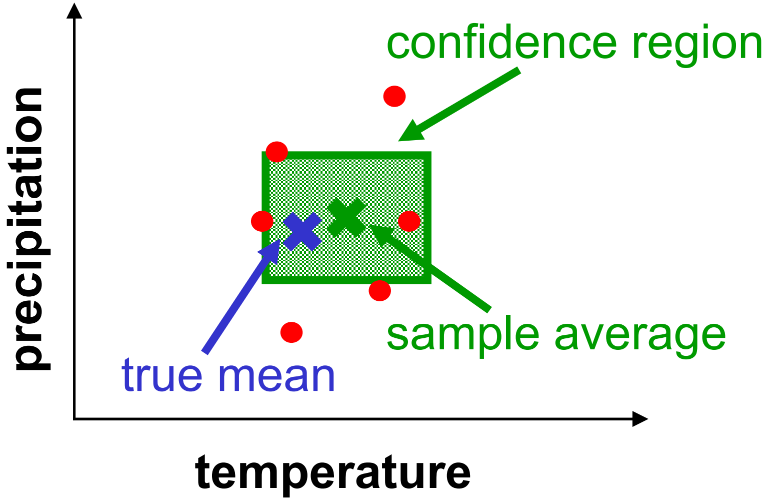
Constraints

Each feature type imposes constraints on output distribution

Linear features	…	mean
Quadratic features	…	variance
Product features	…	covariance
Threshold features	…	proportion above threshold
Hinge features	…	mean above threshold
Binary features (categorical)	…	proportion in each category

Regularization





find distribution of maximum entropy such that
Mean(**f**) in confidence region of sample average of **f**

The Maxent distribution

… is always a Gibbs distribution:

$$q_{\lambda}(x) = exp(\Sigma_j \lambda_j f_j(x)) / Z$$

Z is a scaling factor so distribution sums to 1

f_j is the j'th feature

λ_j is a coefficient, calculated by the program

Maxent is penalized maximum likelihood

Log likelihood:

$$LogLikelihood(q_\lambda) = 1/m \, \Sigma_i \, ln(q_\lambda(x_i))$$

where $x_1 \ldots x_m$ are the occurrence points.

Maxent maximizes regularized likelihood:

$$LogLikelihood(q_\lambda) - \Sigma_j \, \beta_j |\lambda_j|$$

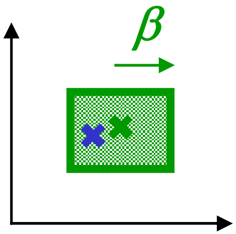
where β_j is the width of the confidence interval for f_j

Similar to Akaike Information Criterion (AIC), lasso.

Performance guarantees

If true mean lies in confidence region then for best Gibbs $\mathbf{q}_\lambda$:

$$\mathrm{RE}(\text{truth} \parallel \mathrm{SOL}) - \mathrm{RE}(\text{truth} \parallel q_{\boldsymbol{\lambda}})$$
$$\leq 2 \cdot \beta \|\boldsymbol{\lambda}\|_1$$



f_j's bounded in $[0,1]$:

$$\beta \propto \sqrt{\frac{\ln n}{m}}$$

f_j's binary of VC-dimension d:

$$\beta \propto \sqrt{\frac{d}{m} \ln \frac{m}{d}}$$

Maxent software: β tuned on a reference data set

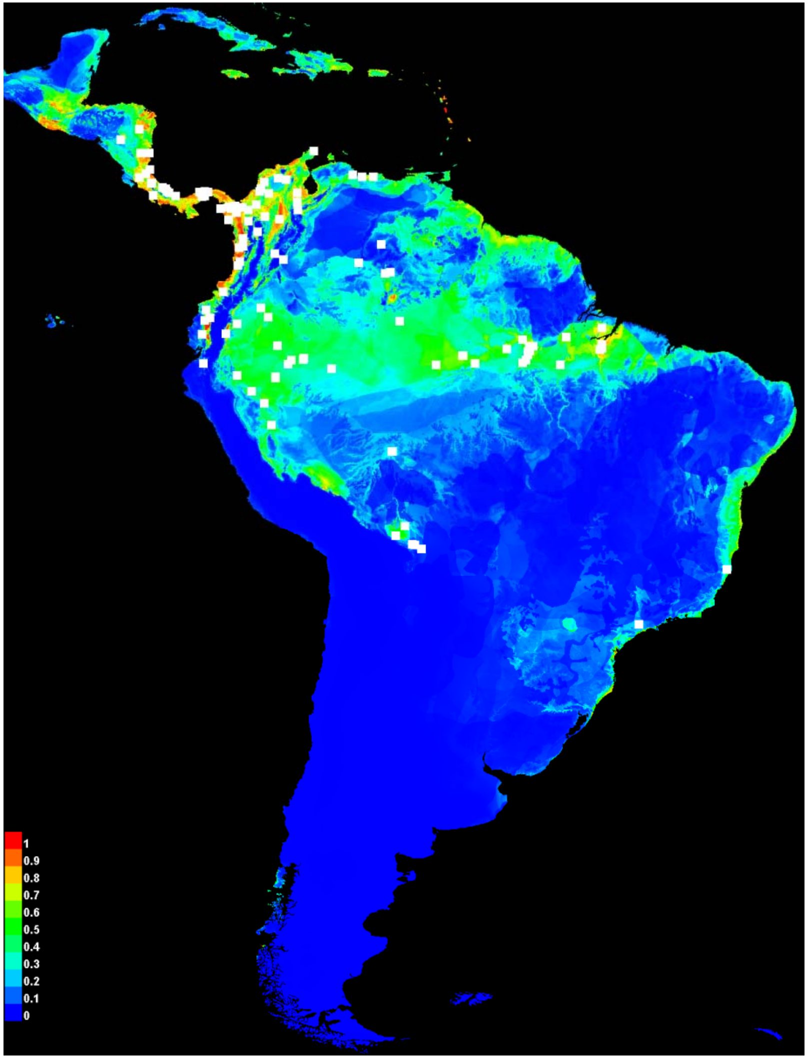
Estimating probability of presence

- Prevalence: Number of sites where the species is present, or sum of probability of presence

- Prevalence not identifiable from occurrence data (Ward et al. 2009)
  - Example: sparrow and sparrow-hawk
  - Both have same range map
  - Both have same geographic distribution of occurrences
  - Hawk is rarer within its range: lower prevalence

- Probability of presence & prevalence depend on sampling:
  - Site size
  - Observation time

Logistic output format



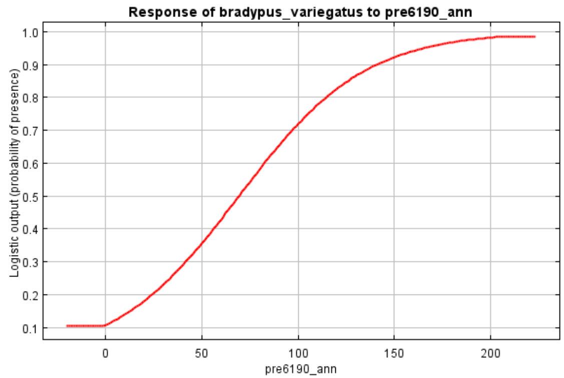
- **Minimax: maximize performance for worst-case prevalence**
- **Exponential → logistic model**
  - Offset term: entropy
- **Scaled so “typical” presences have value 0.5**

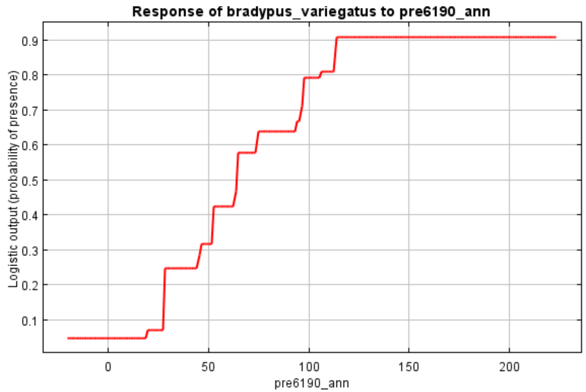
Response curves

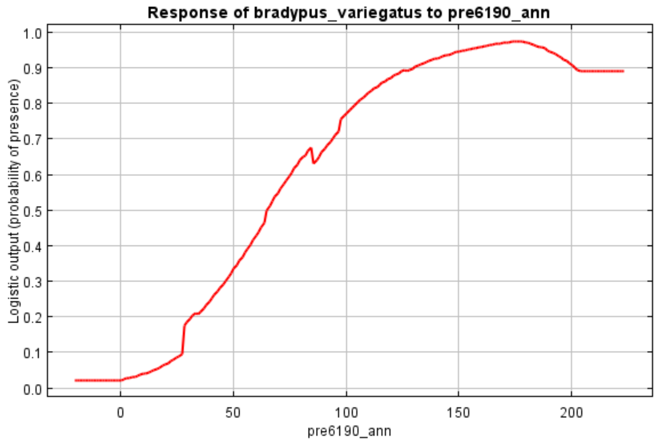
- How does probability of presence depend on each variable?

- Simple features → simpler model
- Complex features → complex model

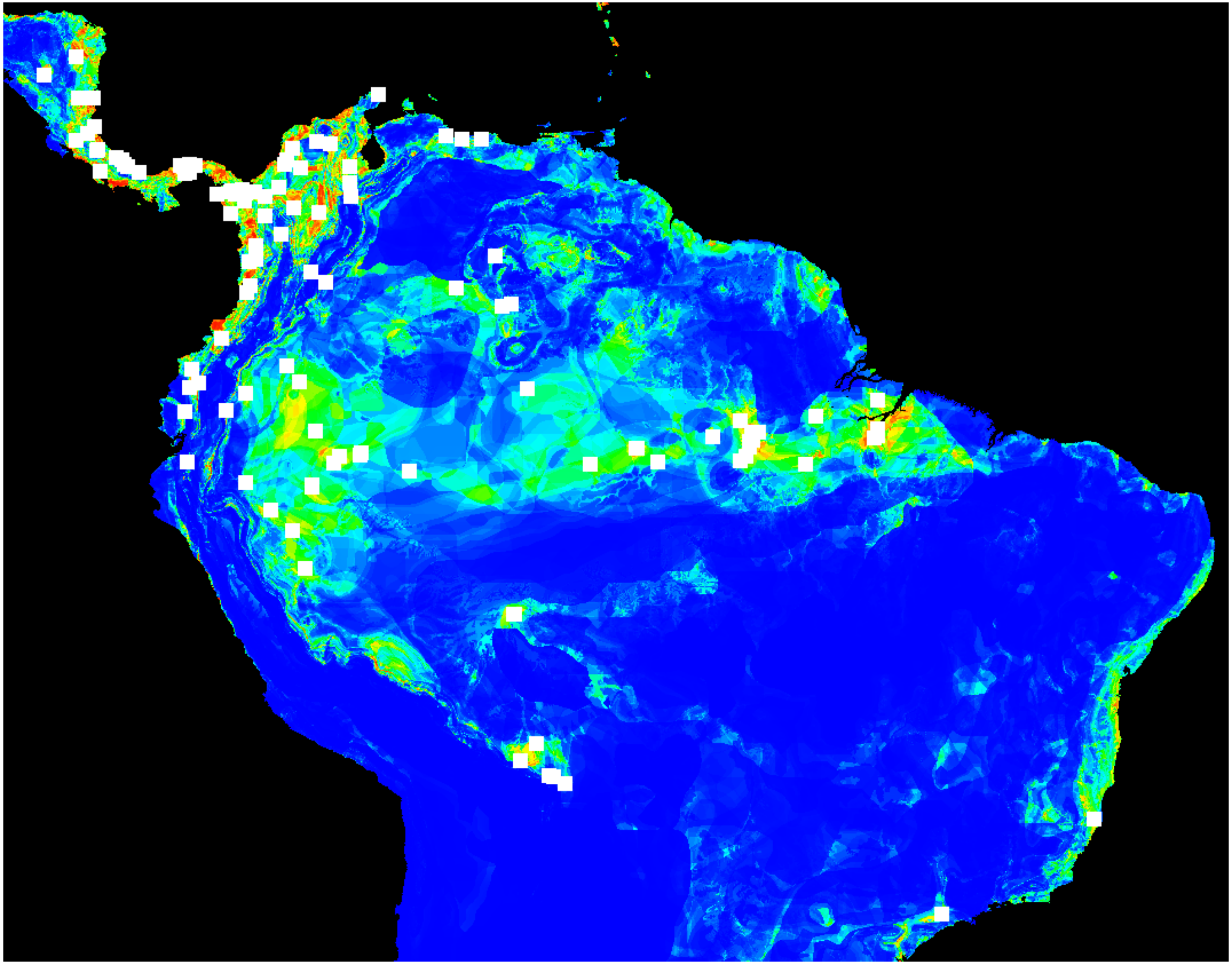
- Linear + quadratic (top)
- Threshold features (middle)
- All feature types (bottom)







Effect of regularization: multiplier = 0.2

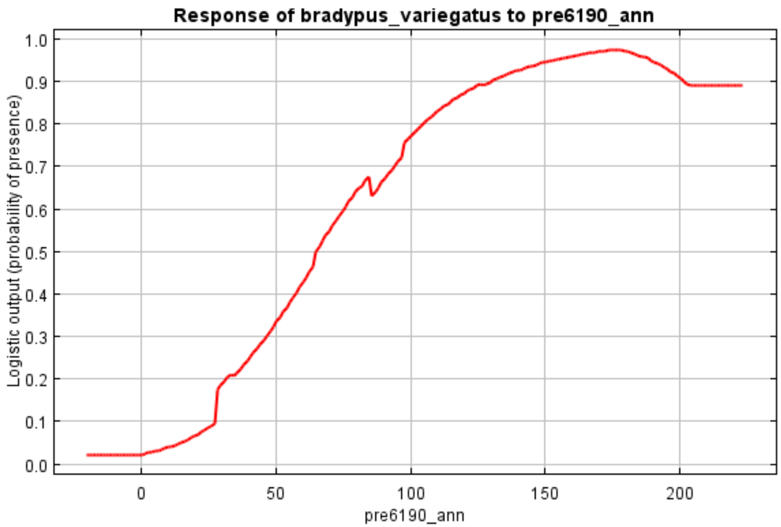


Smaller confidence Intervals

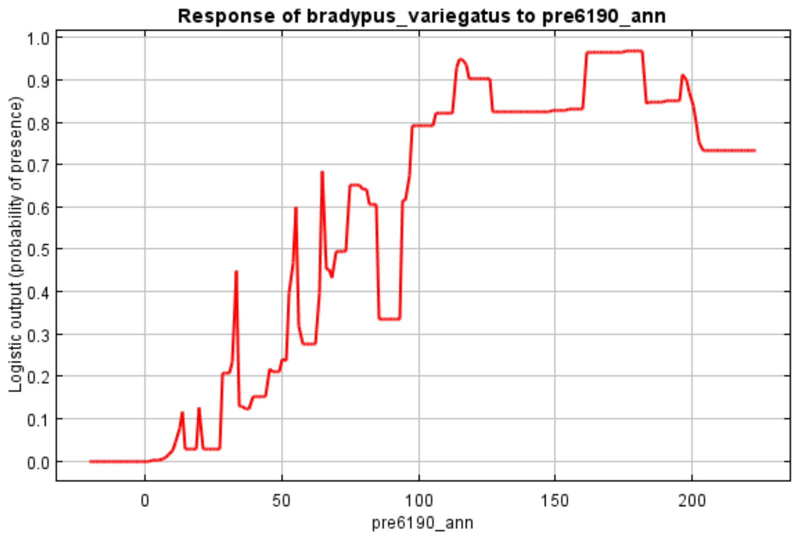
Lower entropy

Less spread-out

Effect of regularization: over-fitting



Regularization multiplier = 1.0
(not over-fit)

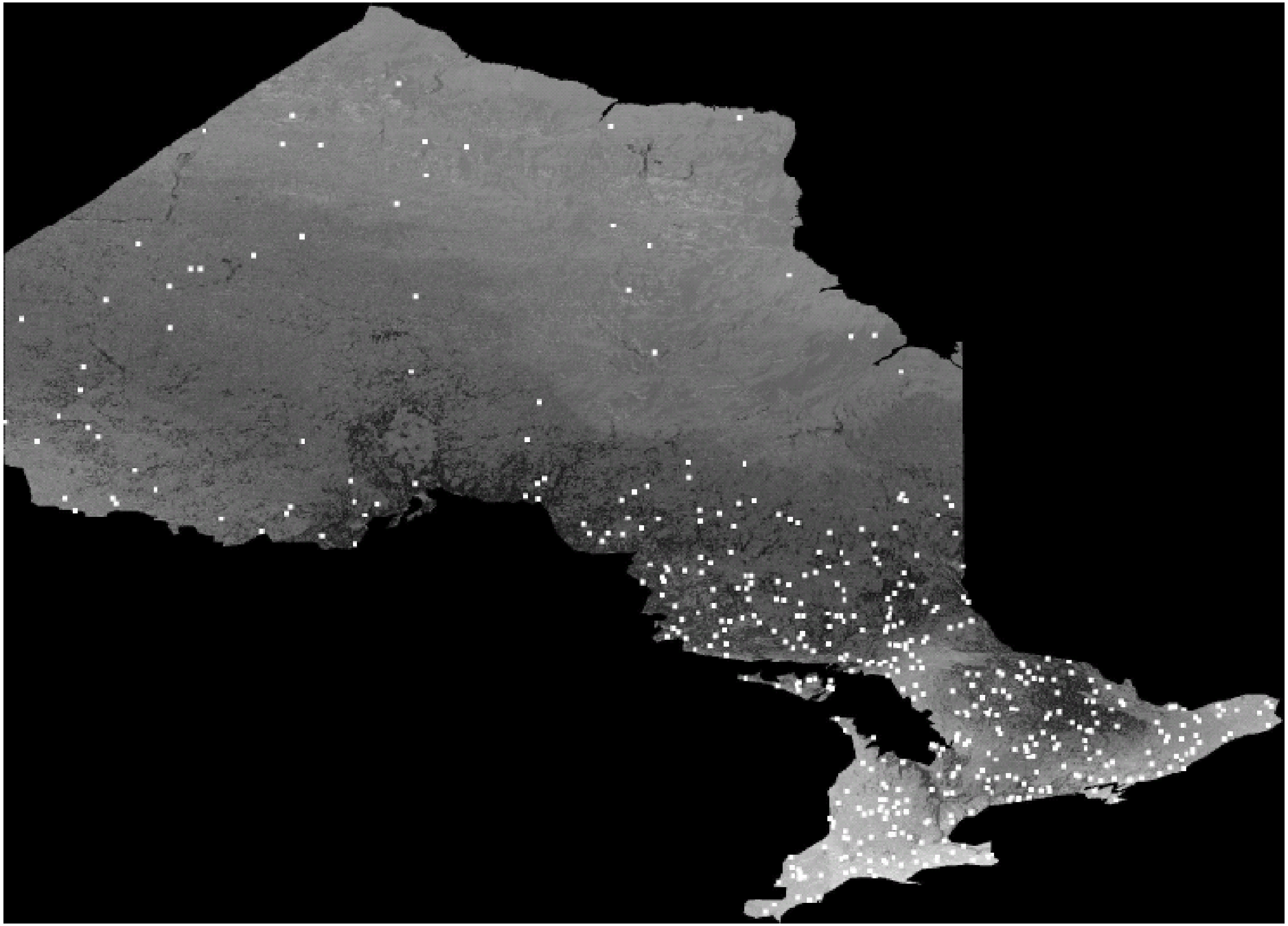


Regularization multiplier = 0.2
(clearly over-fit)

The dangers of bias

- Virtual species in Ontario, Canada
  - prefers mid-range of all climatic variables

Boosted regression tree model: biased p/a data

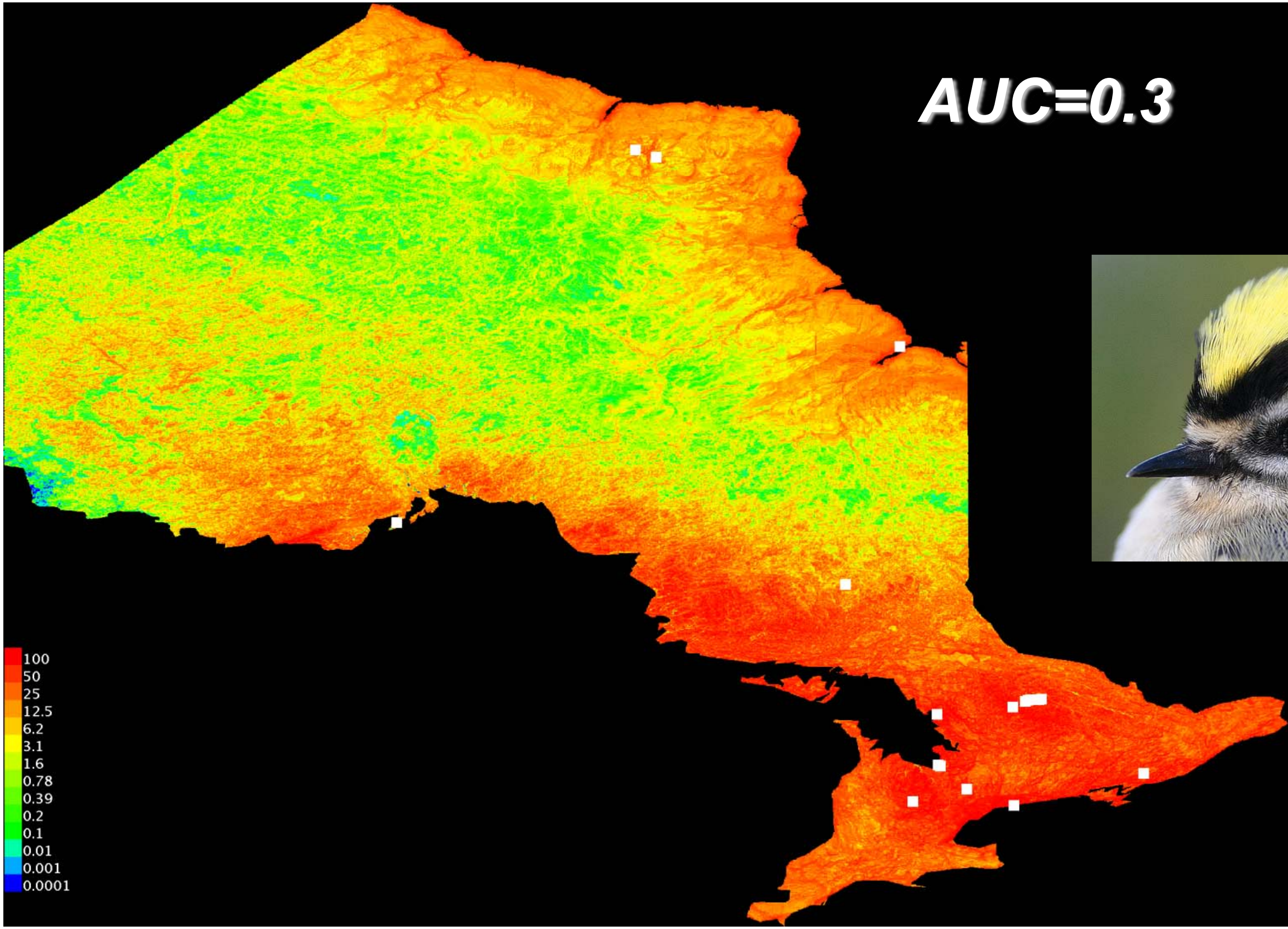


Presence-absence model recovers species distribution

Model from biased occurrence data

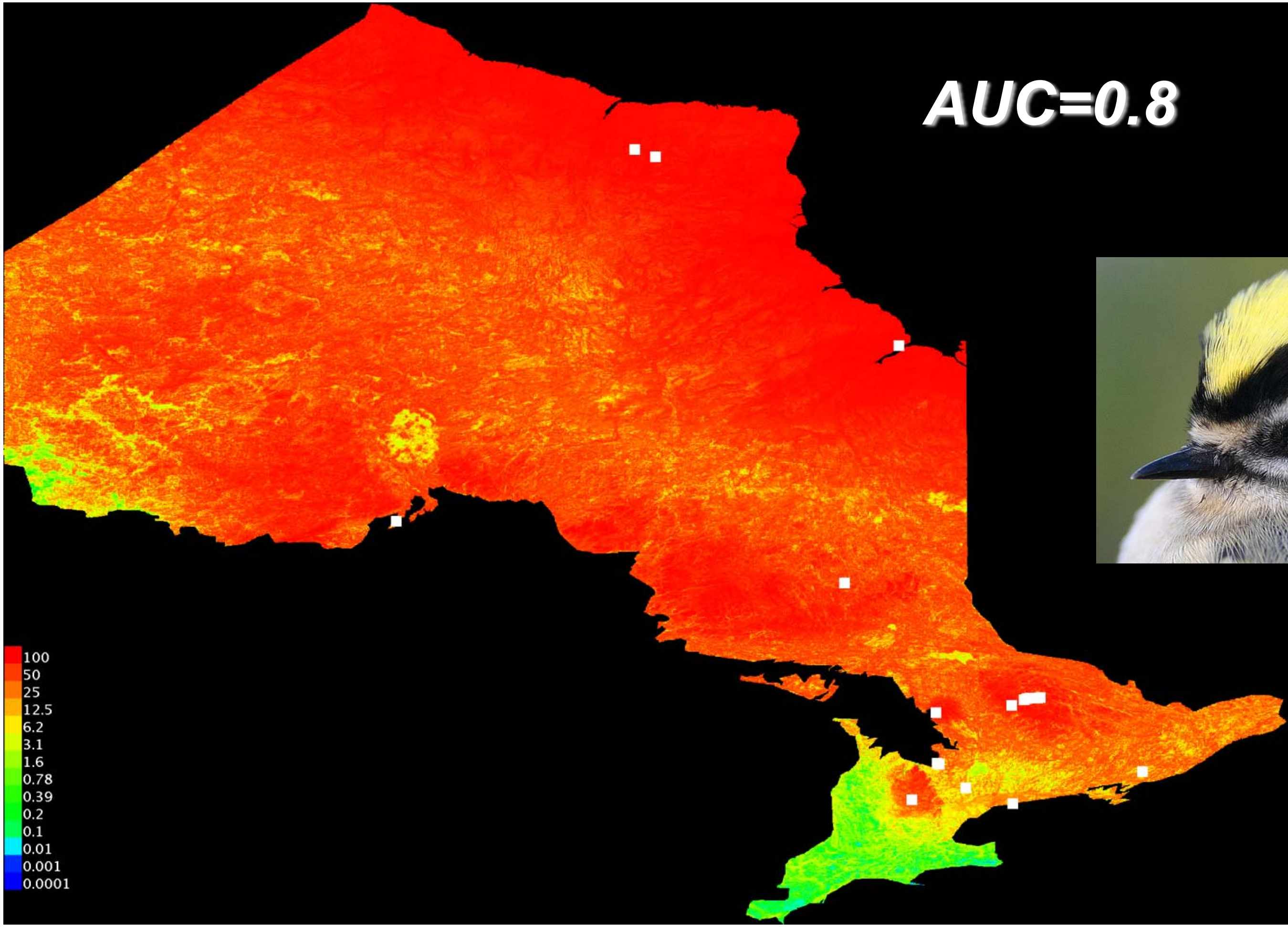
Model recovers sampling bias, not species distribution

Correcting bias: golden-crowned kinglet



Maxent model from biased occurrence data

Correcting bias with target-group background



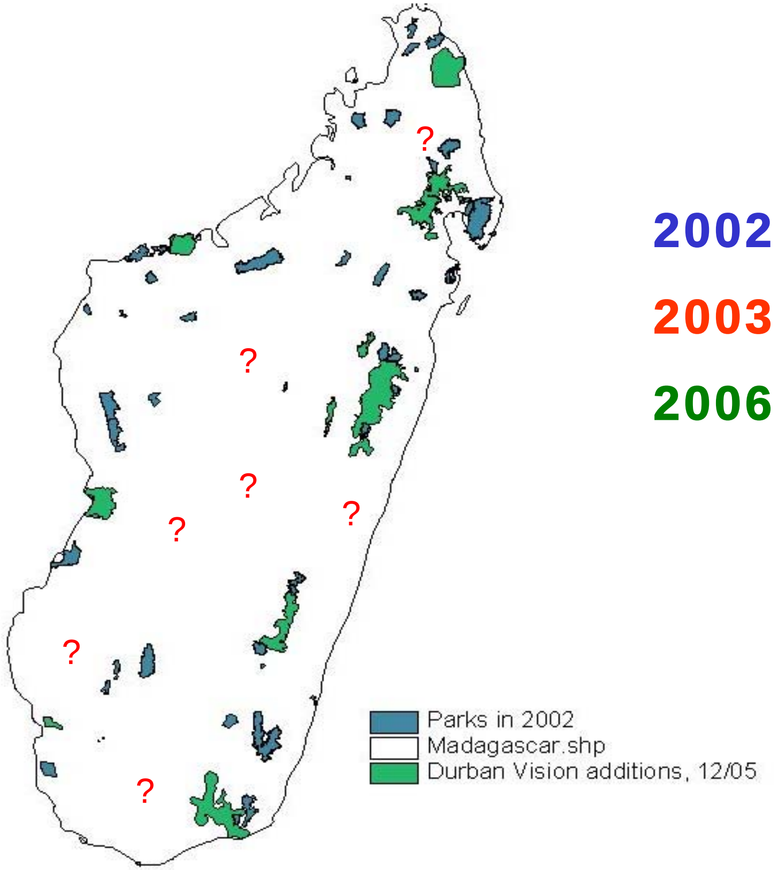
- **Infer sampling distribution from other species' records**
  - **"Target group", collected by same methods**

Aligning Conservation Priorities Across Taxa in Madagascar with High-Resolution Planning Tools

C. Kremen, A. Cameron *et al.*

Science **320**, 222 (2008)

Madagascar: Opportunity Knocks



2002: 1.7 million ha = 2.9%

2003 Durban Vision: 6 million ha = 10%

2006: 3.88 million ha = 6.3%

Study outline

- **Gather biodiversity data**
  - **2315 species: lemurs, frogs, geckos, ants, butterflies, plants**
  - **Presences only, limited data, sampling biases**

- **Model species distributions:** Maxent

- **New reserve selection software**: Zonation
  - **1 km2 resolution for entire country**
  - **> 700,000 units**

Mystrium mysticum, dracula ant

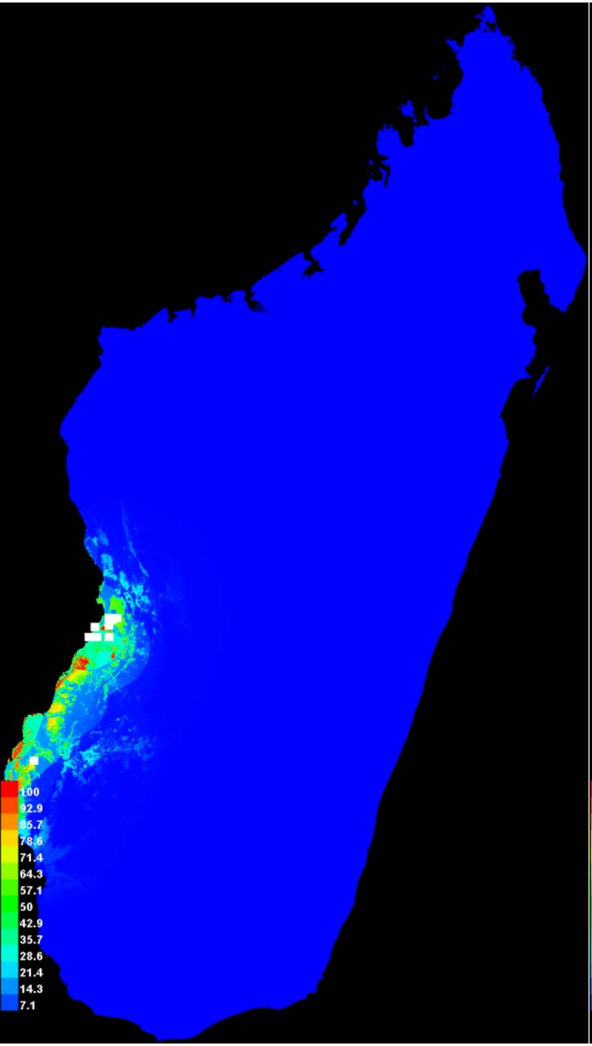
Adansonia grandidieri, Grandidier's baobab

Uroplatus fimbriatus, common leaf-tailed gecko

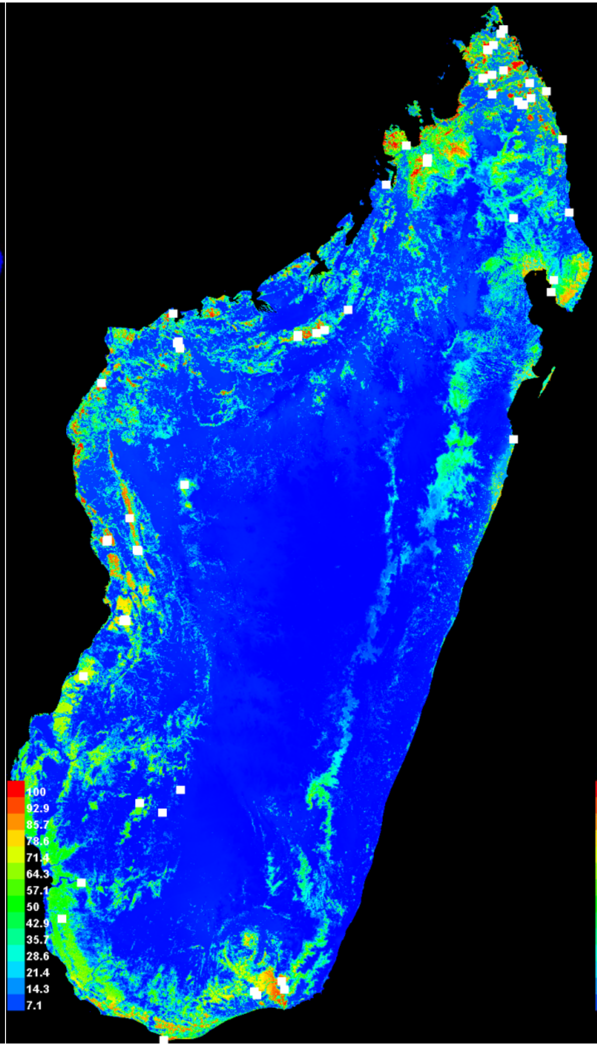
Indri indri

Propithecus diadema, diademed sifaka

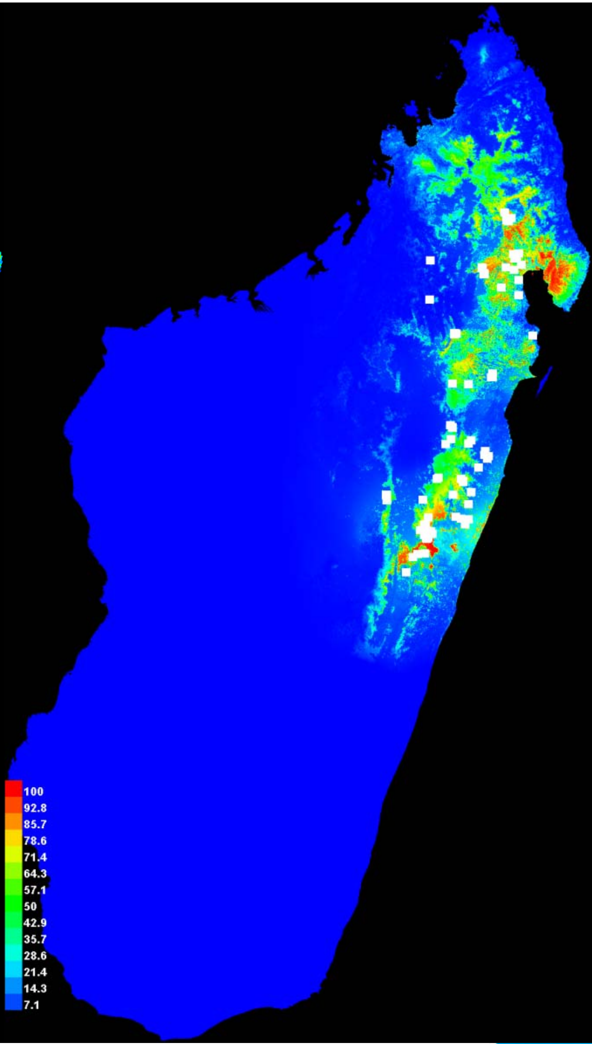
Grandidier’s baobab

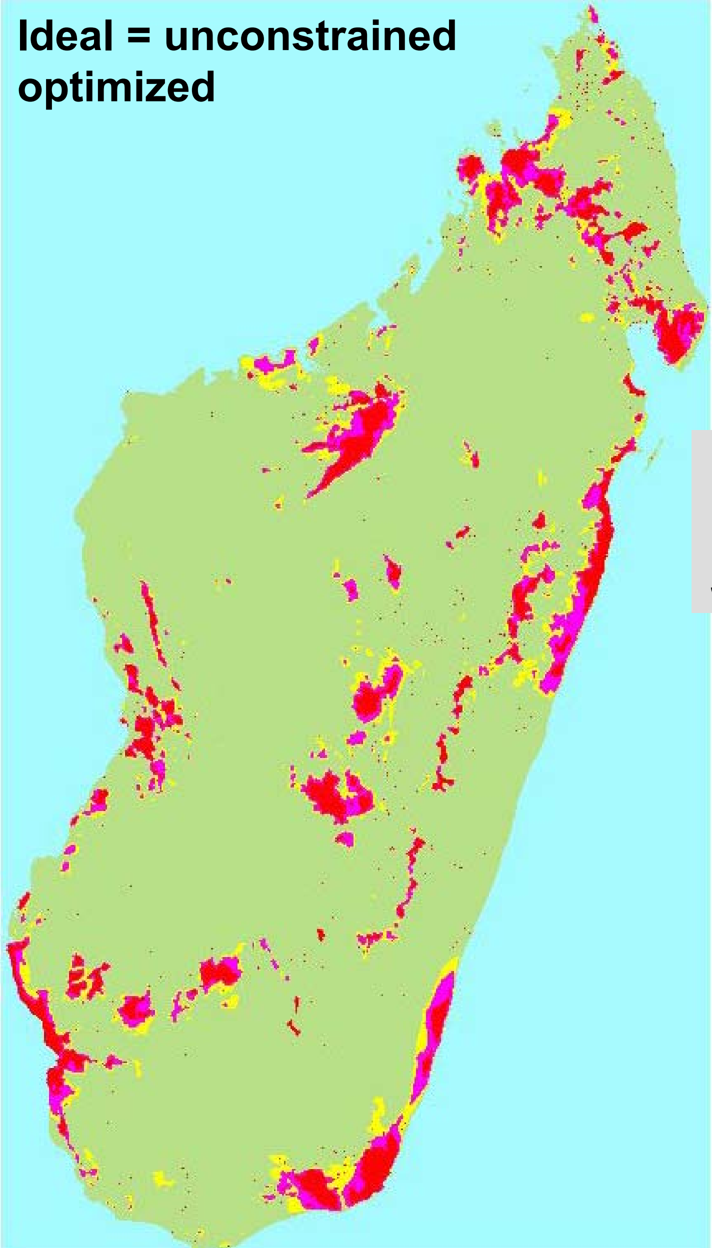


Dracula ant

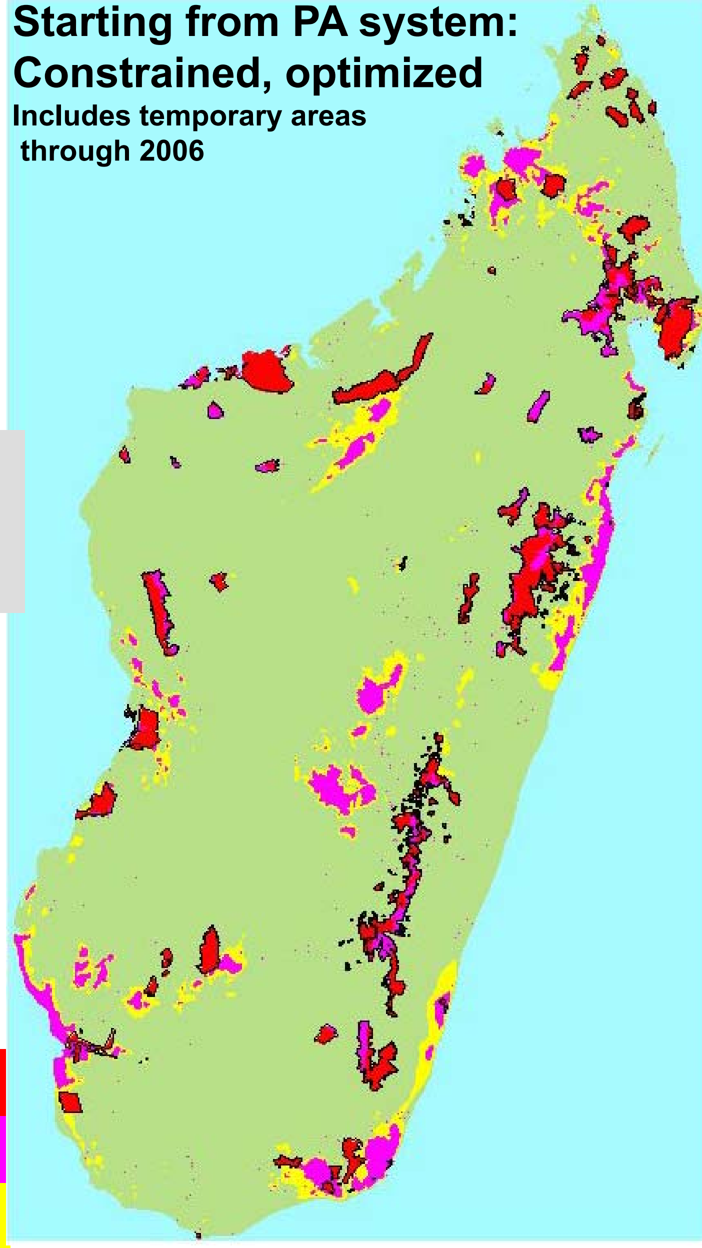


Indri





Multi-taxon Solutions



TOP 5%

5 to 10%

10 to 15%

Spare slides

Maximum Entropy Principle

The fact that a certain probability distribution maximizes entropy subject to certain constraints representing our incomplete information, is the fundamental property which justifies the use of that distribution for inference; it agrees with everything that is known but carefully avoids assuming anything that is not known (Jaynes, 1990).

Maximizing "gain"

Unregularized gain:

$$Gain(q_\lambda) = Log\ likelihood - ln(1/n)$$

E.g. if UGain=1.5, then average training sample is exp(1.5) (about 4.5) times more likely than a random background pixel

Maxent maximizes regularized gain:

$$Gain(q\boldsymbol{\lambda}) - \Sigma_j\, \beta_j|\lambda_j|$$

Maxent algorithms

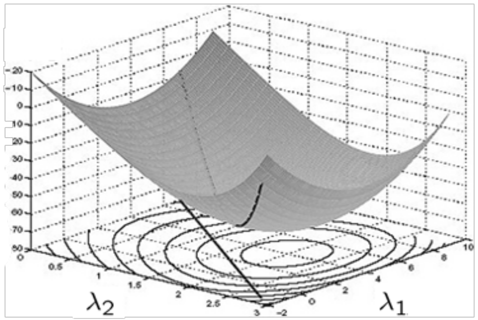
Goal: maximize the regularized gain

Algorithm:

Start with uniform distribution (gain=0)

Iteratively update **λ** to increase the gain

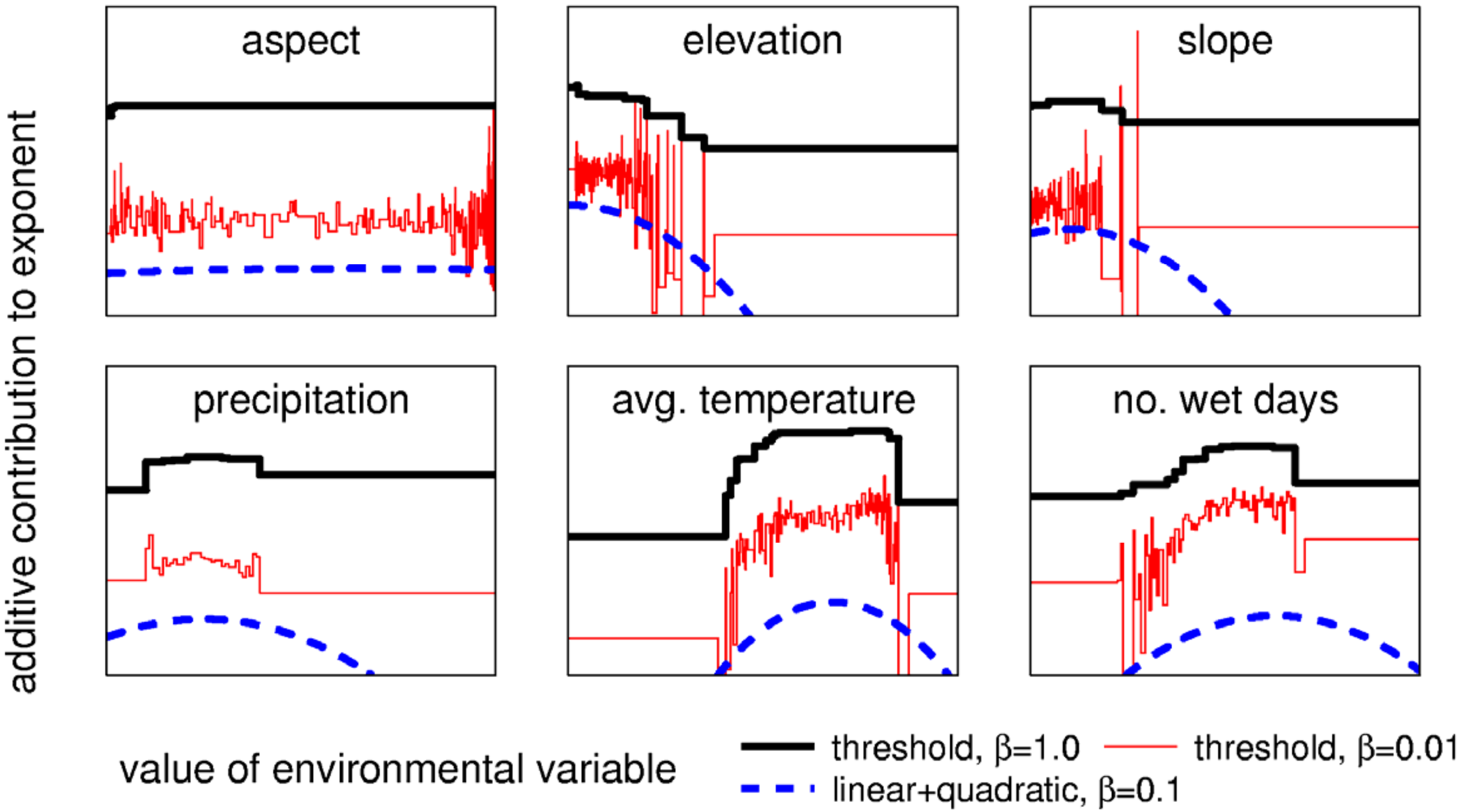
Regularized gain (λ_1, λ_2)



The gain is convex:

- **Variety of algorithms: gradient descent, conjugate gradient, Newton, iterative scaling**
- **Our algorithm: coordinate descent**

Interpretation of regularization



Conditional vs unconditional Maxent

One class:

- Distribution over sites: $p(x|y=1)$
- Maximize entropy: $-\Sigma p(x|y=1) \ln(p(x|y=1))$

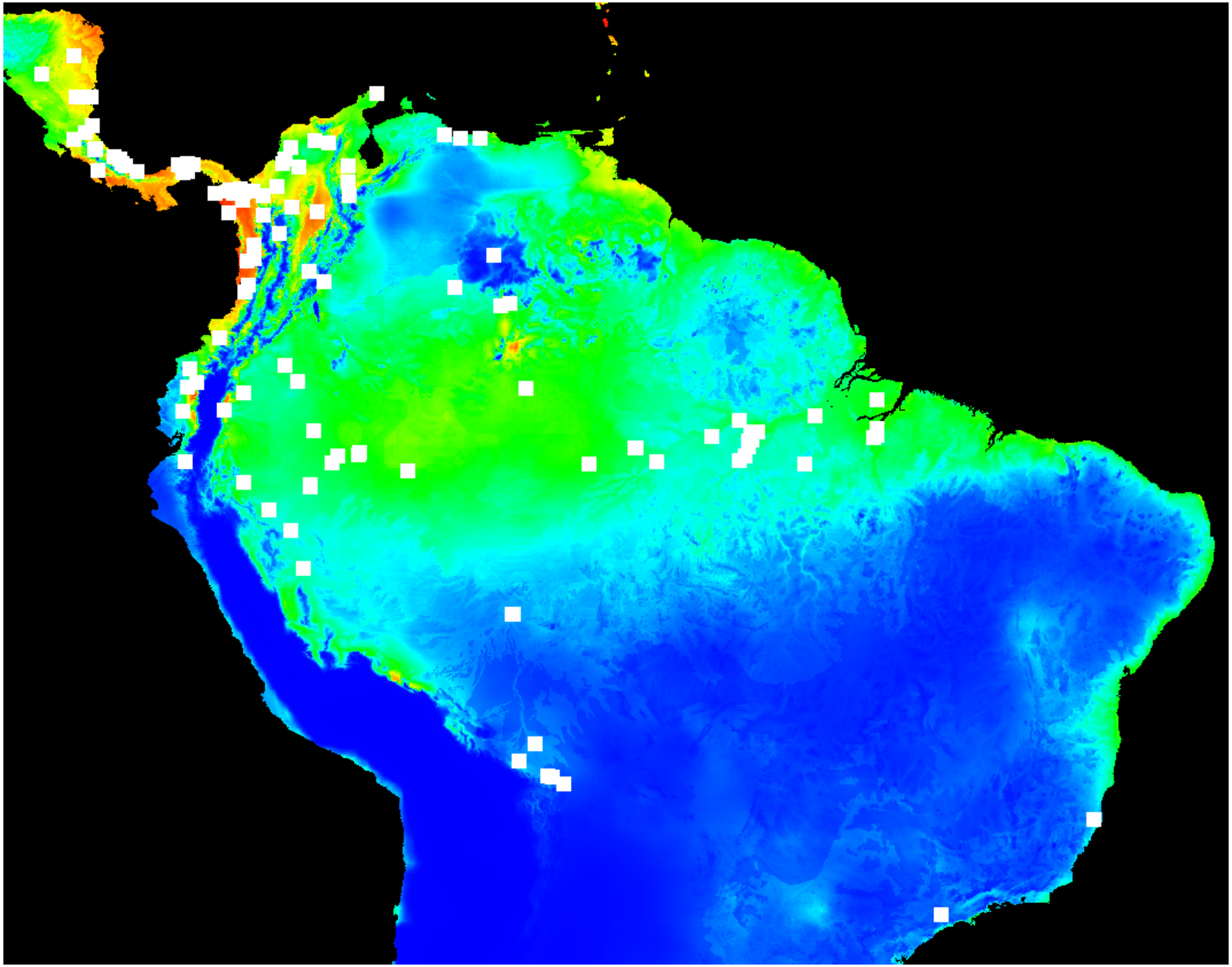
Multiclass:

- Conditional probability of presence: $Pr(y|\mathbf{z})$
- Maximize conditional entropy: $-\Sigma p'(\mathbf{z})\, p(y|\mathbf{z}) \ln(p(y|\mathbf{z}))$

Notation:

- y 0 or 1, species presence
- x a site in our study region
- $\mathbf{z}$ a vector of environmental conditions
- $p'(z)$ the empirical probability of $\mathbf{z}$

Effect of regularization: multiplier = 5

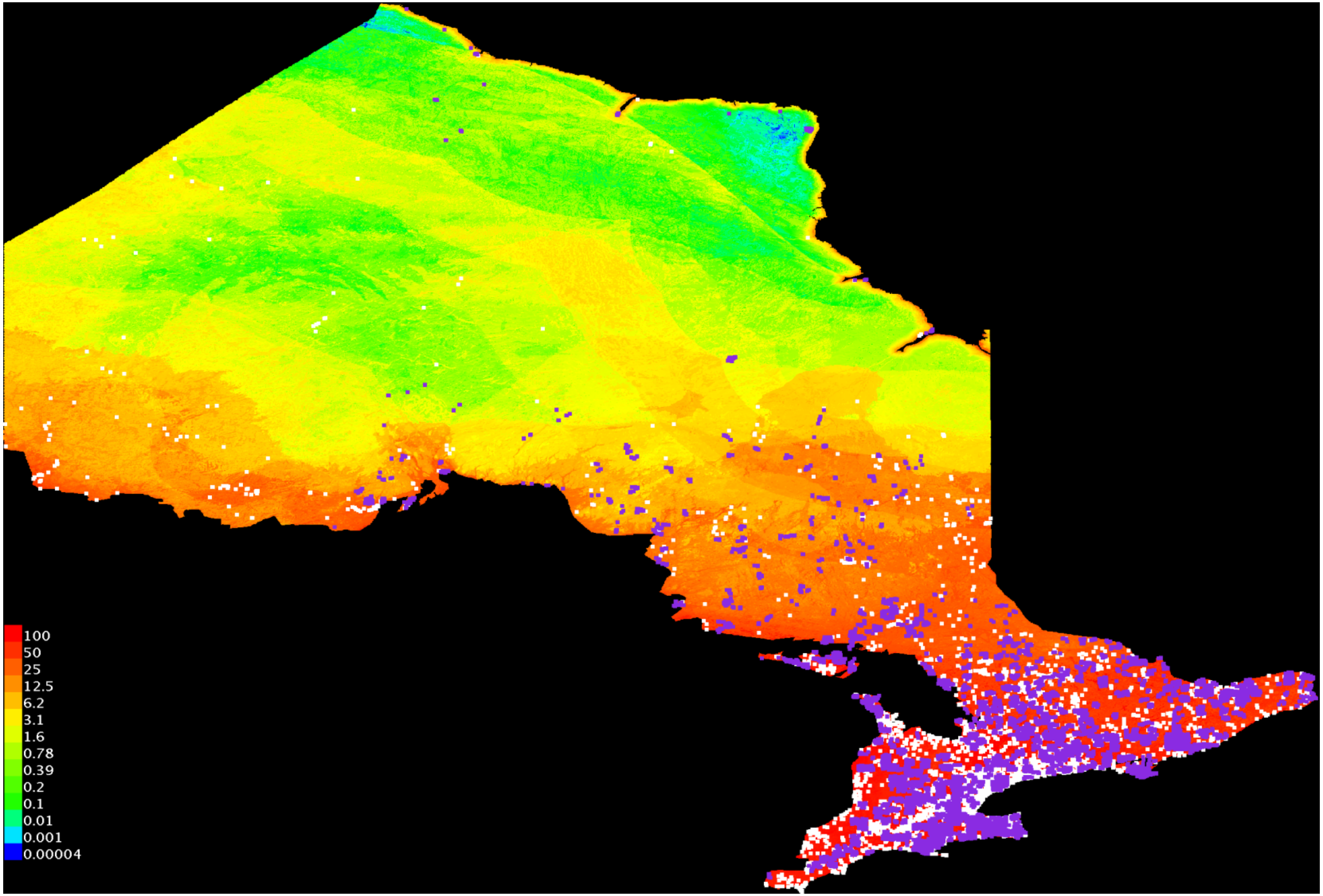


Larger confidence Intervals

Higher entropy

More spread-out

Sample selection bias in Ontario birds



Performance guarantees

Solution SOL returned by Maxent is almost as good as the best $\mathbf{q}_{\boldsymbol{\lambda}}$

$$\mathsf{RE}(\mathsf{truth} \parallel \mathsf{SOL}) - \mathsf{RE}(\mathsf{truth} \parallel q_{\boldsymbol{\lambda}}) \leq \mathsf{small}$$

relative entropy
(KL divergence)

Guarantees should depend on

- number of samples **m**
- number of features **n** (or "complexity" of features)
- "complexity" of the best $\mathbf{q}_{\boldsymbol{\lambda}}$