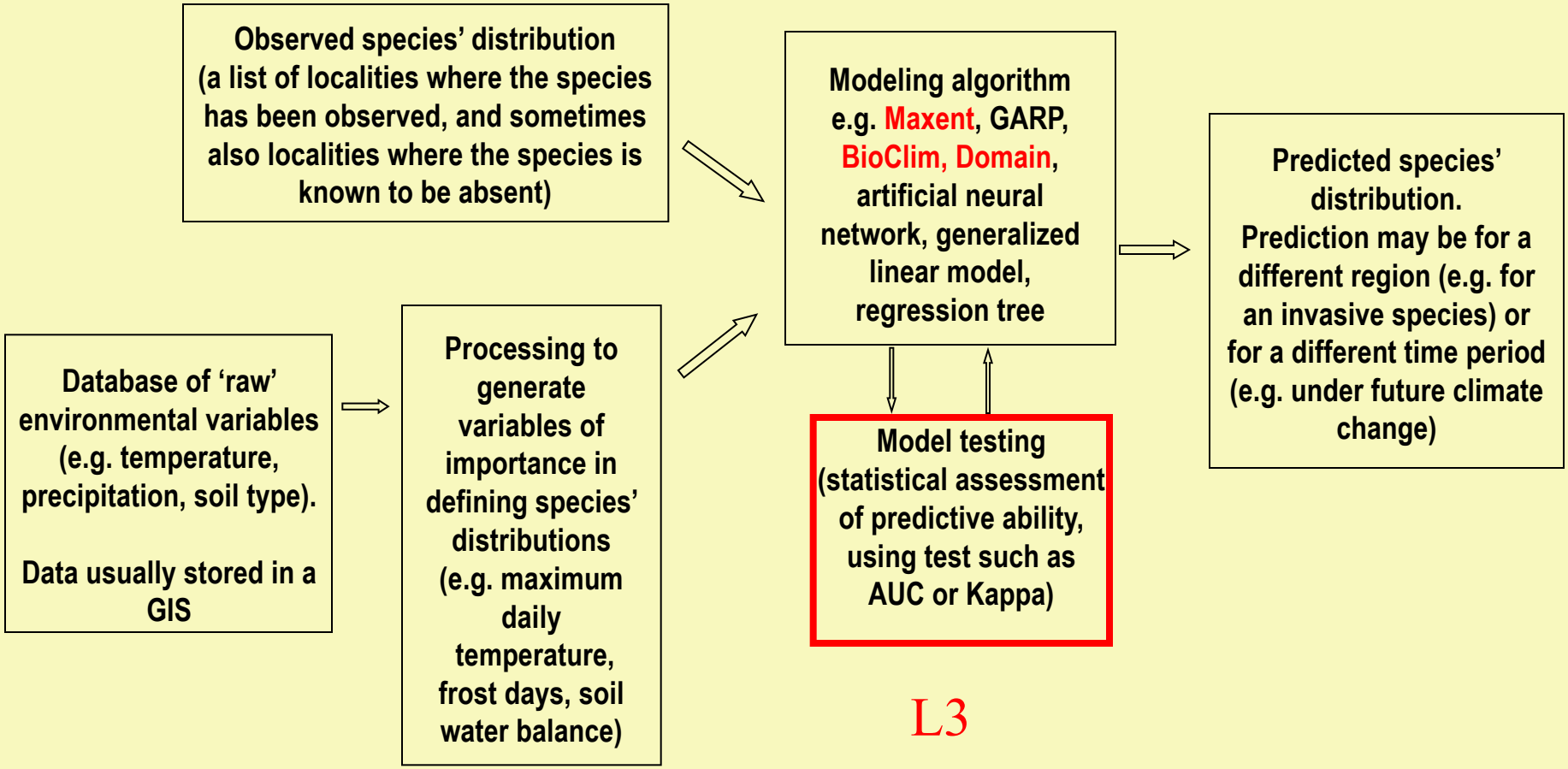
Lecture 4:
Testing predictive performance of niche-based distribution models

Prepared by: Richard Pearson
Adapted by: Alison Cameron

Outline:

- Sources of evaluation data
- Presence-only testing
- Presence-absence testing
- Setting decision thresholds
- Threshold-independent testing

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



L3

Model validation

- Fielding, A. H., and J. F. Bell. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. 24:38-49.
- Austin, M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*. 200:1-19

YouTube Tutorials

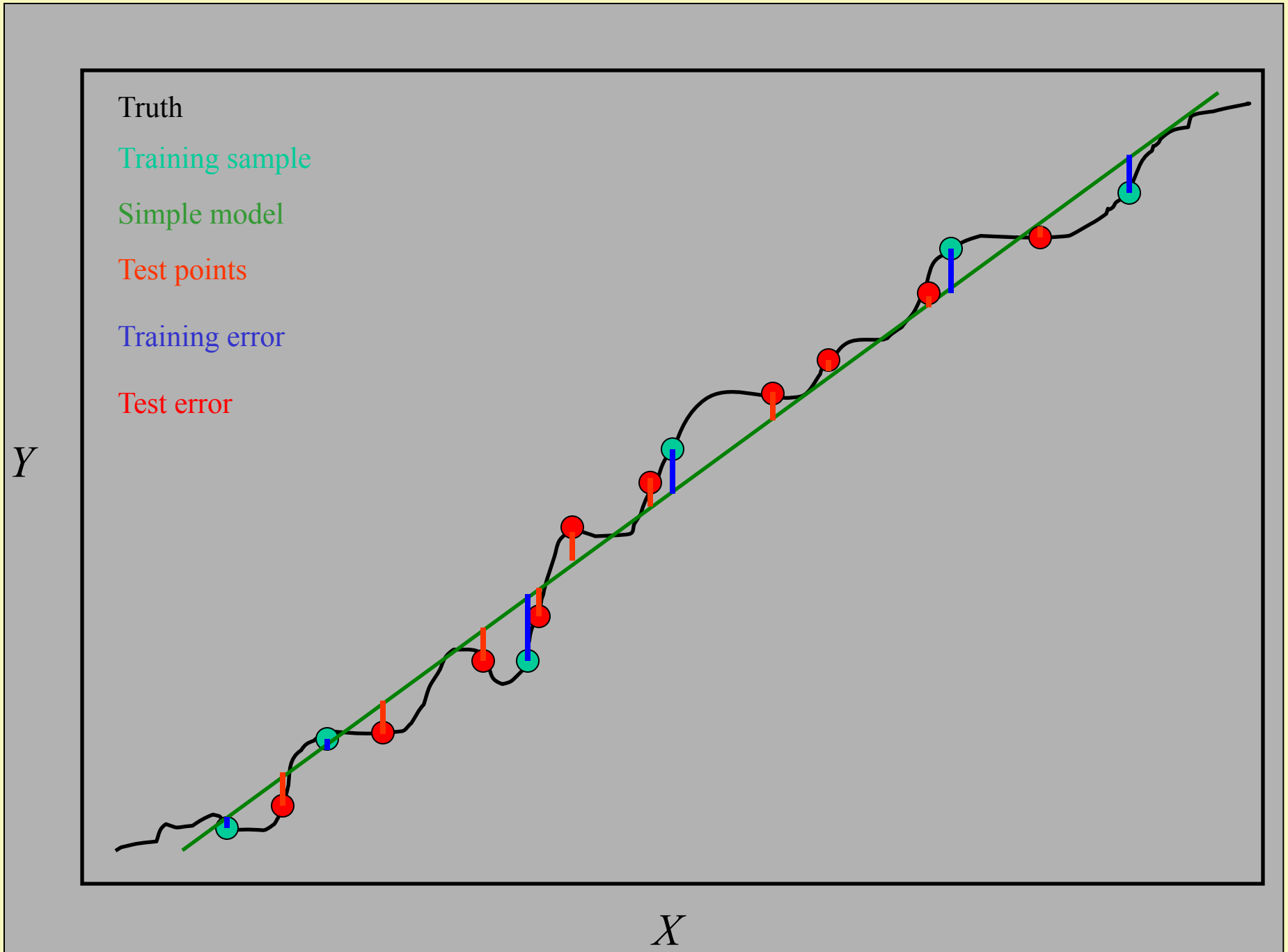
Townsend Peterson

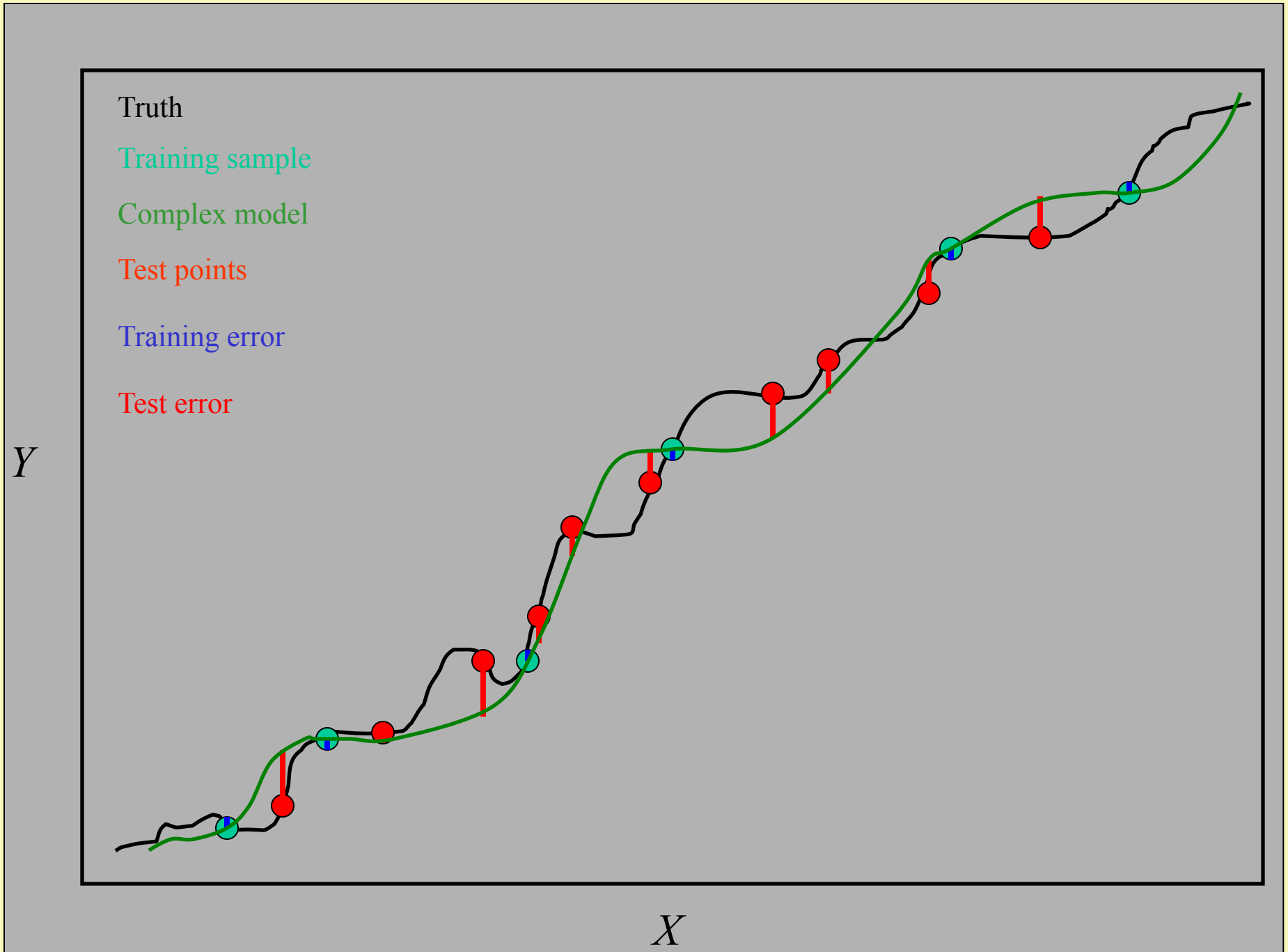
- English

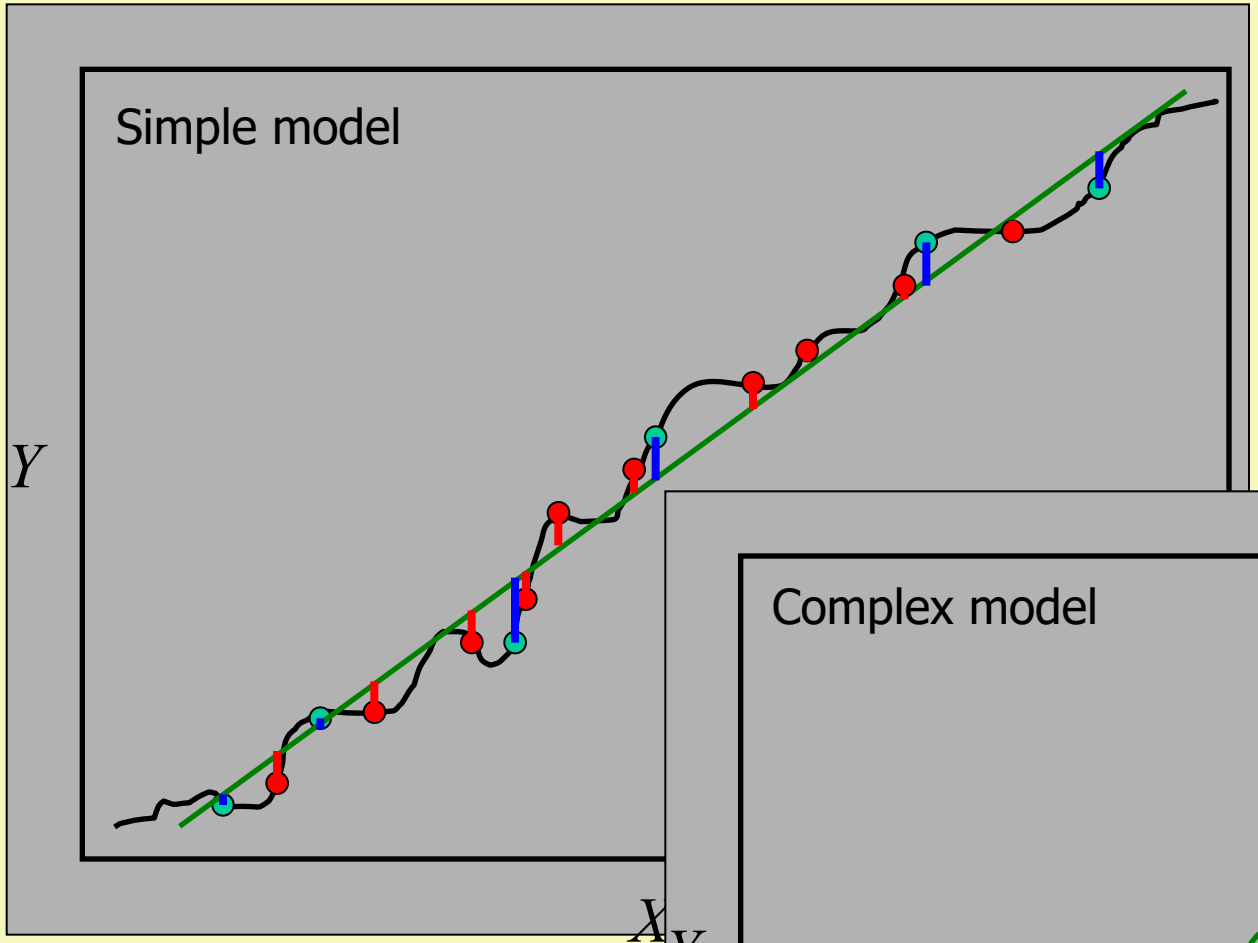
https://www.youtube.com/watch?v=nTLP8oPcPl8

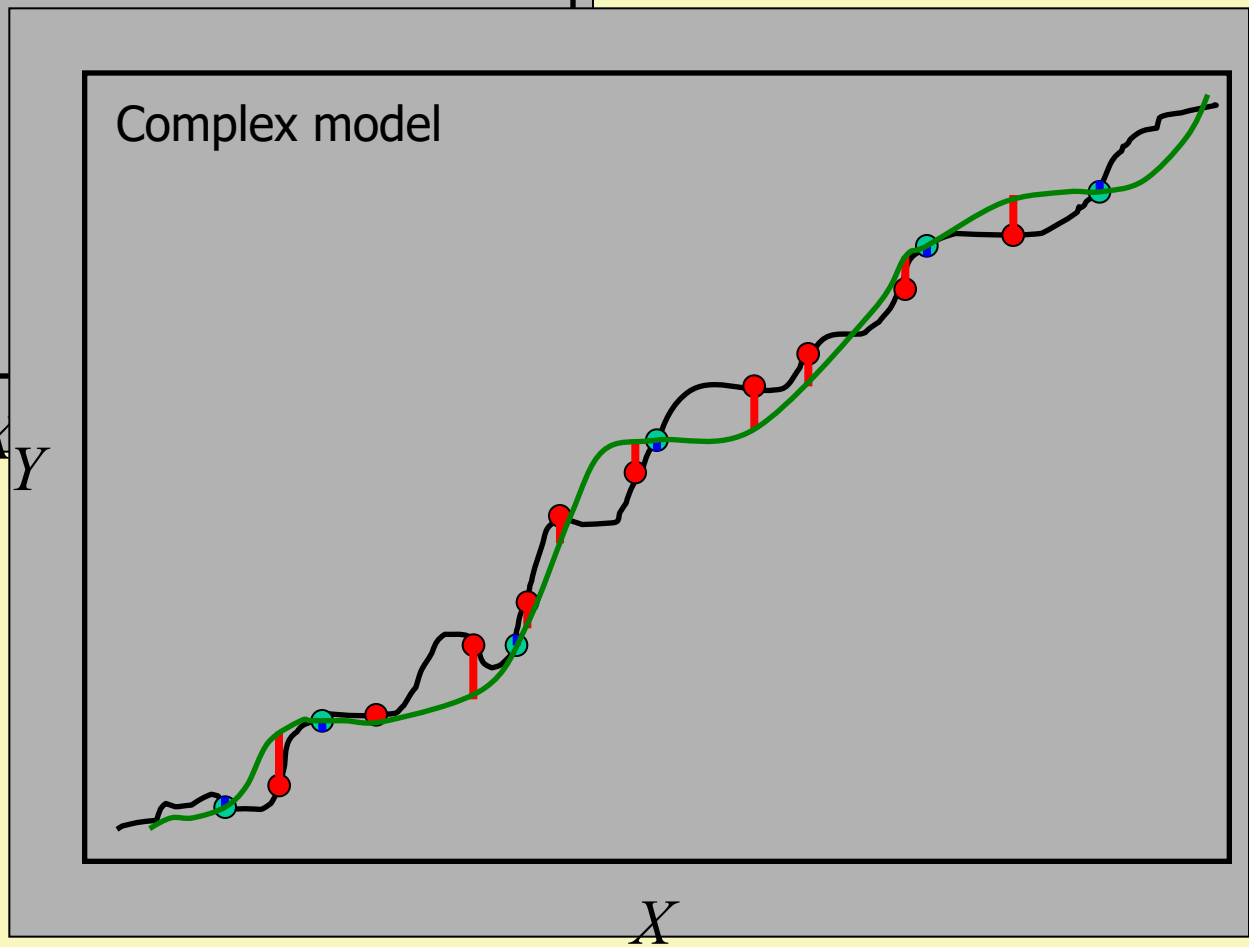
- Portuguese

https://www.youtube.com/watch?v=WCI-hVP-Nt4

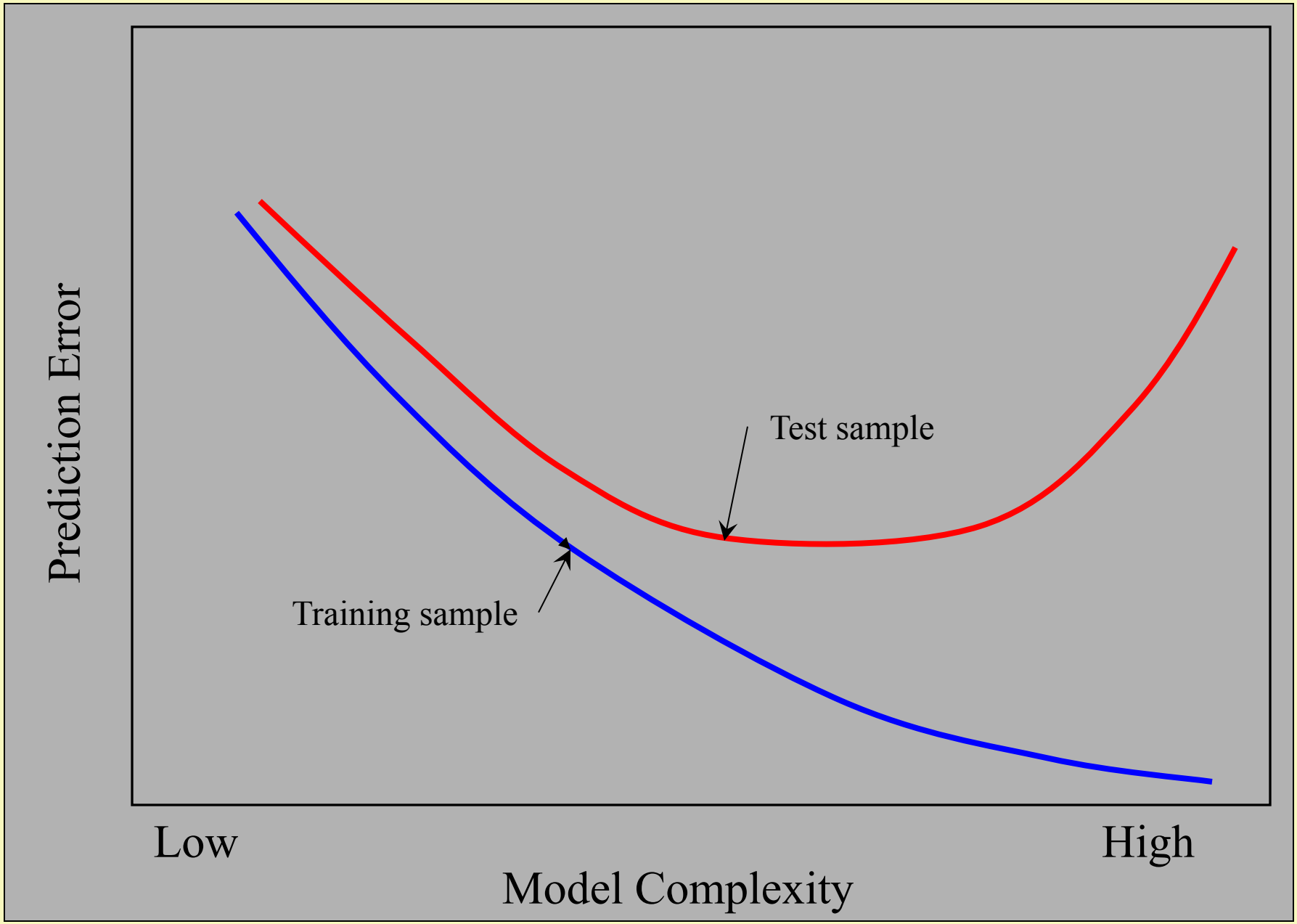

Truth
Training sample
Simple model
Test points
Training error
Test error
Y
X


Truth
Training sample
Complex model
Test points
Training error
Test error
Y
X



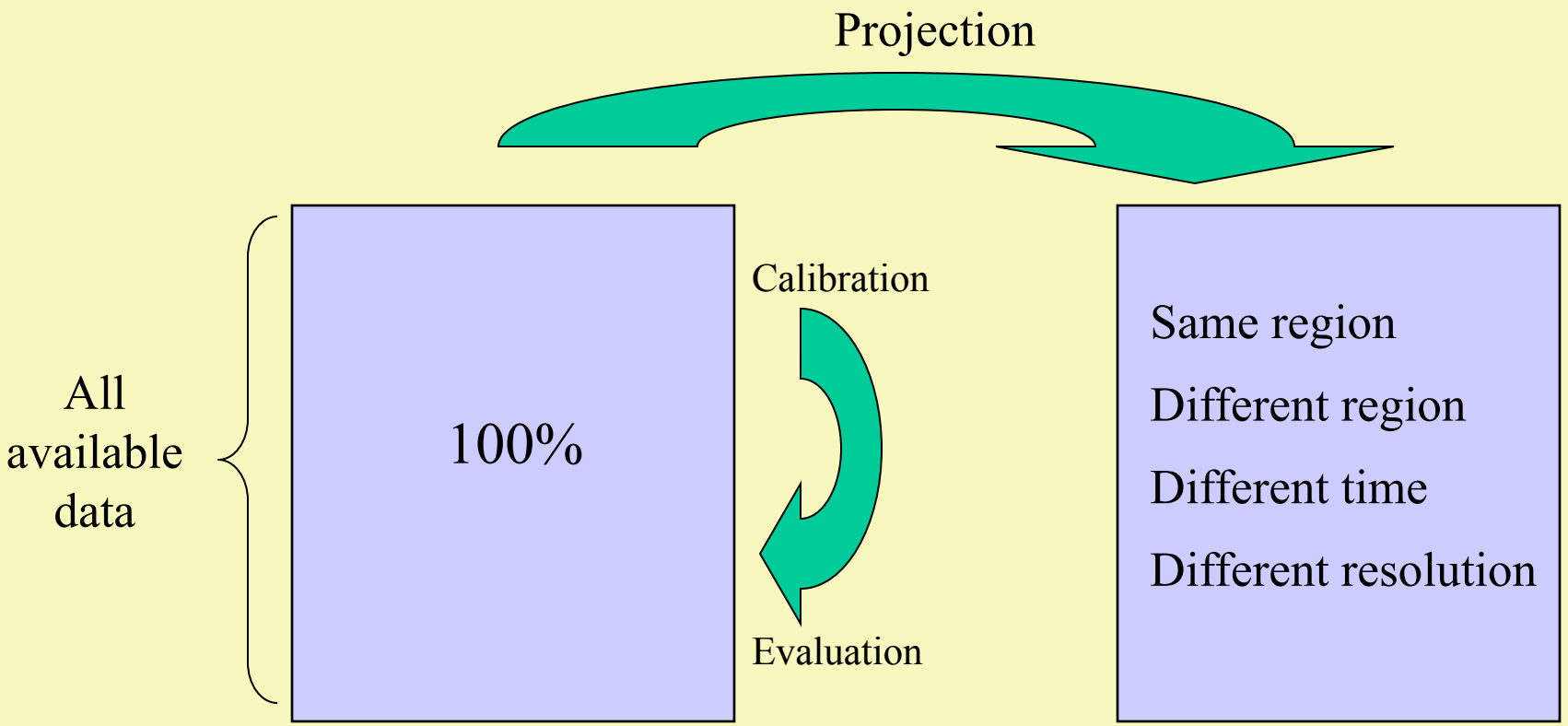


Comparisons



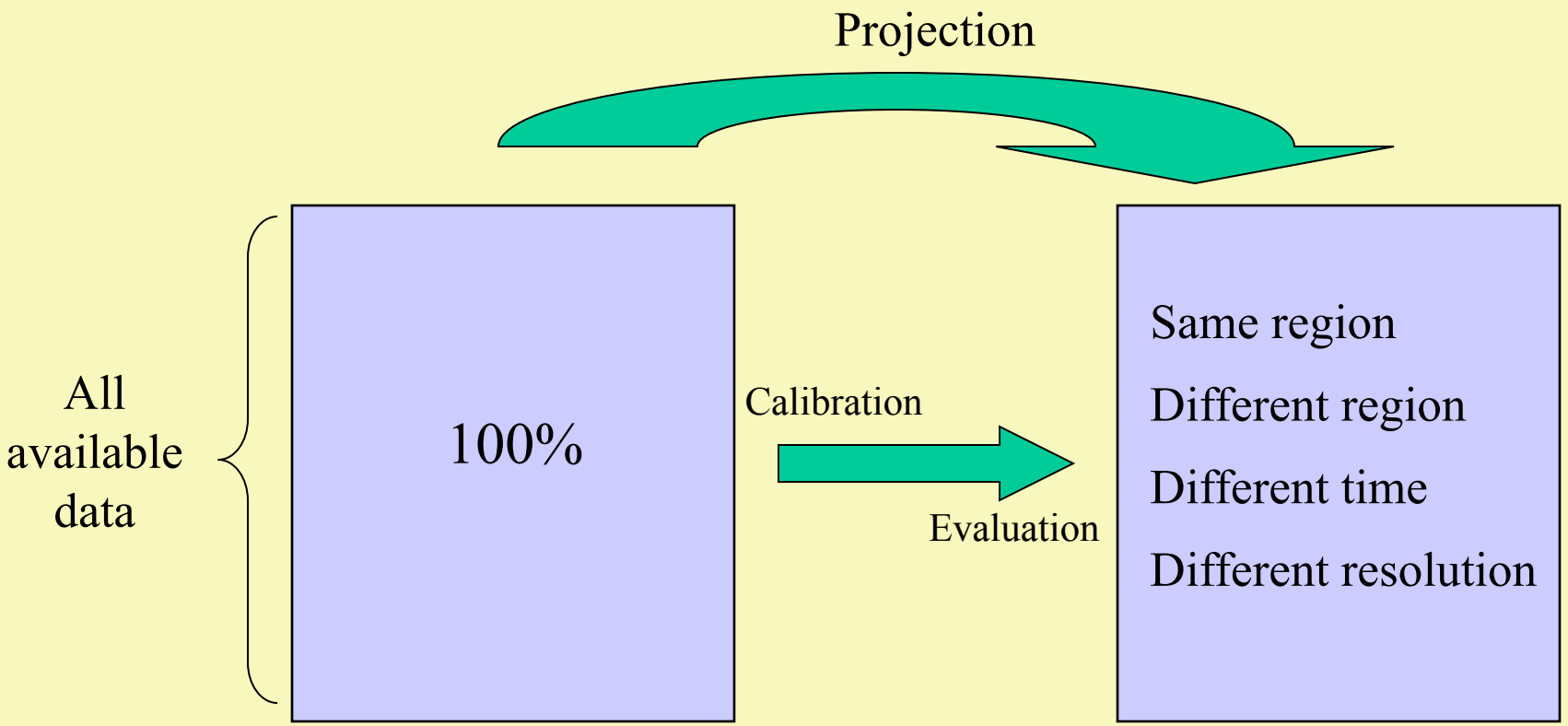
Hastie *et al*. (2001)

Model calibration and evaluation strategies: resubstitution



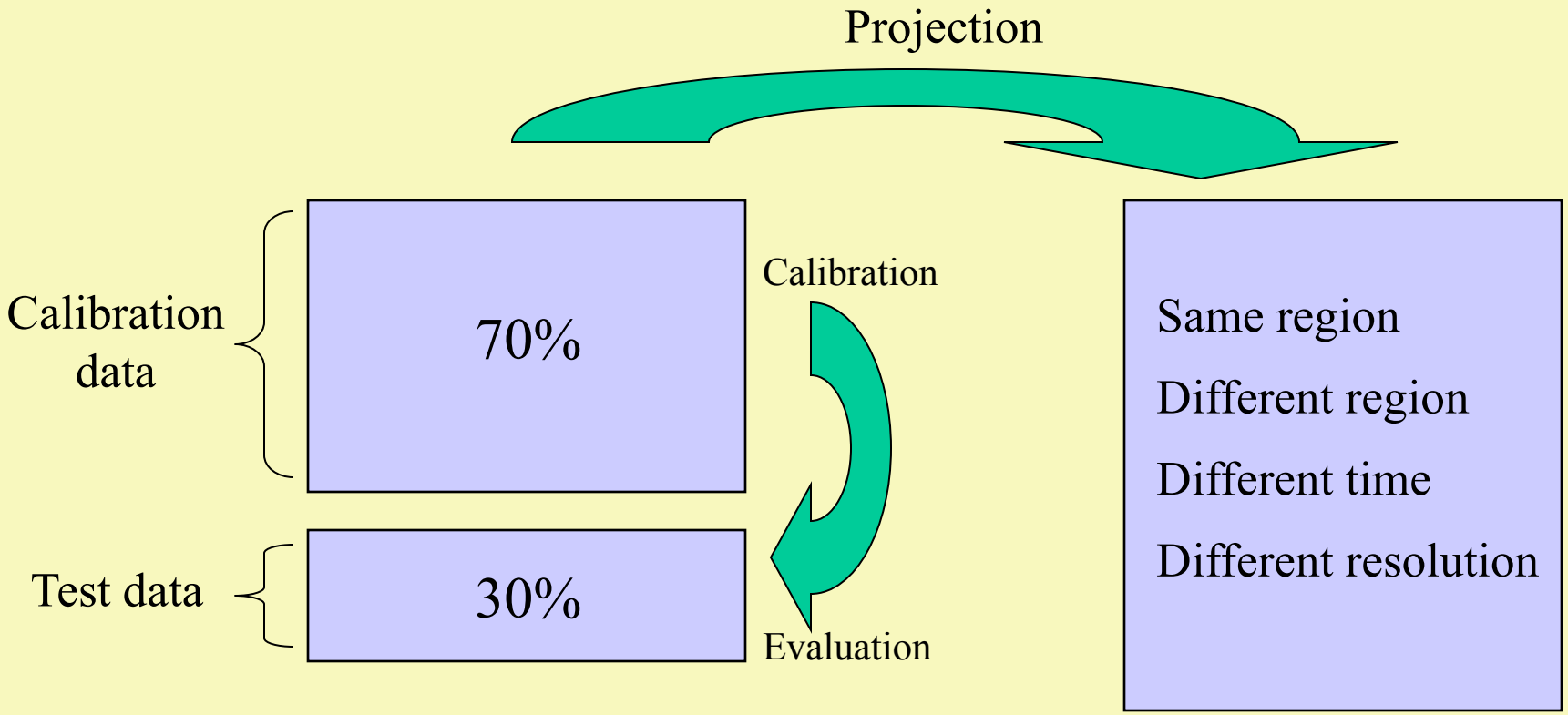
(after Araújo et al. 2005 *Gl. Ch. Biol.*)

Model calibration and evaluation strategies: independent validation



(after Araújo et al. 2005 *Gl. Ch. Biol.*)

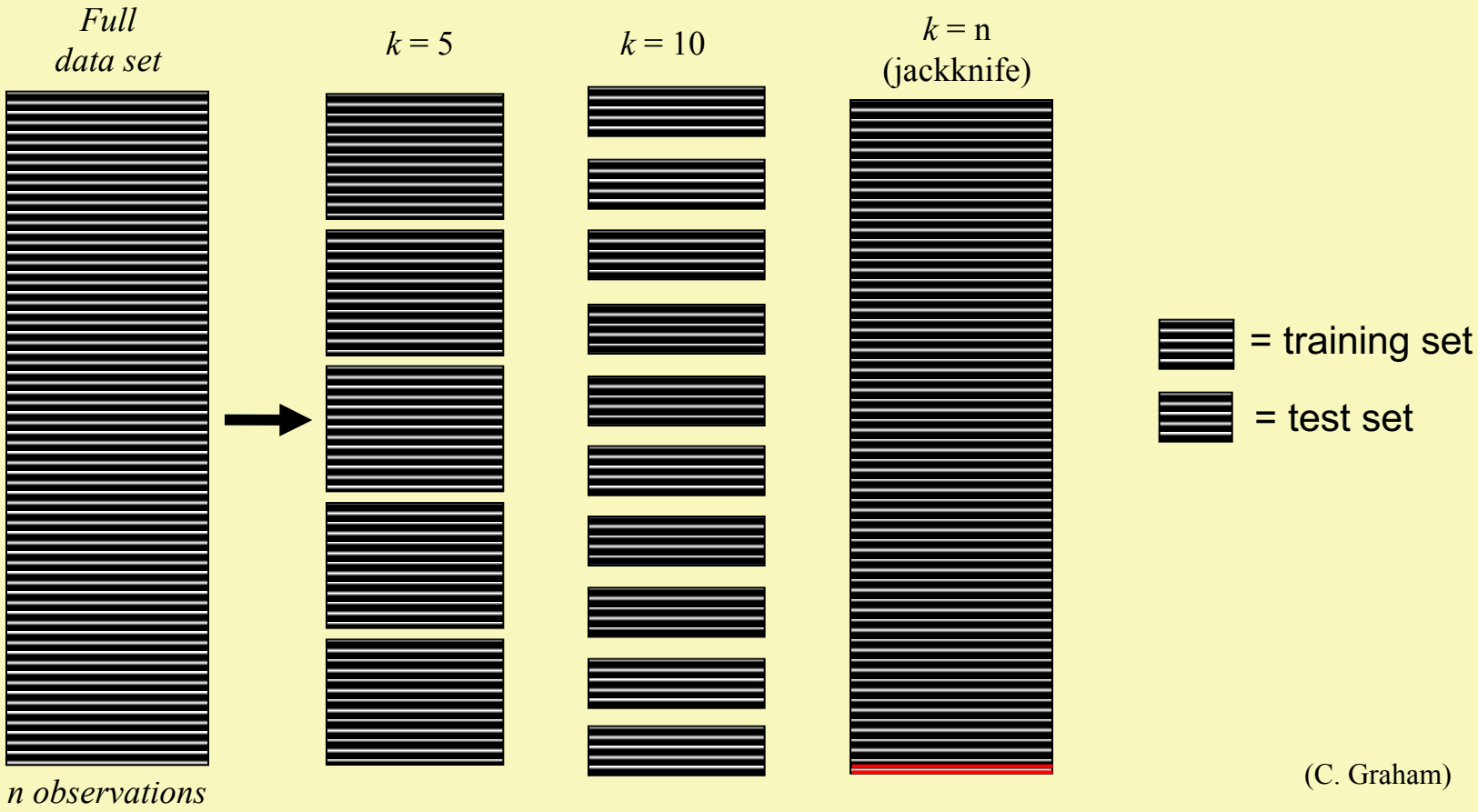
Model calibration and evaluation strategies: data splitting



(after Araújo et al. 2005 *Gl. Ch. Biol.*)

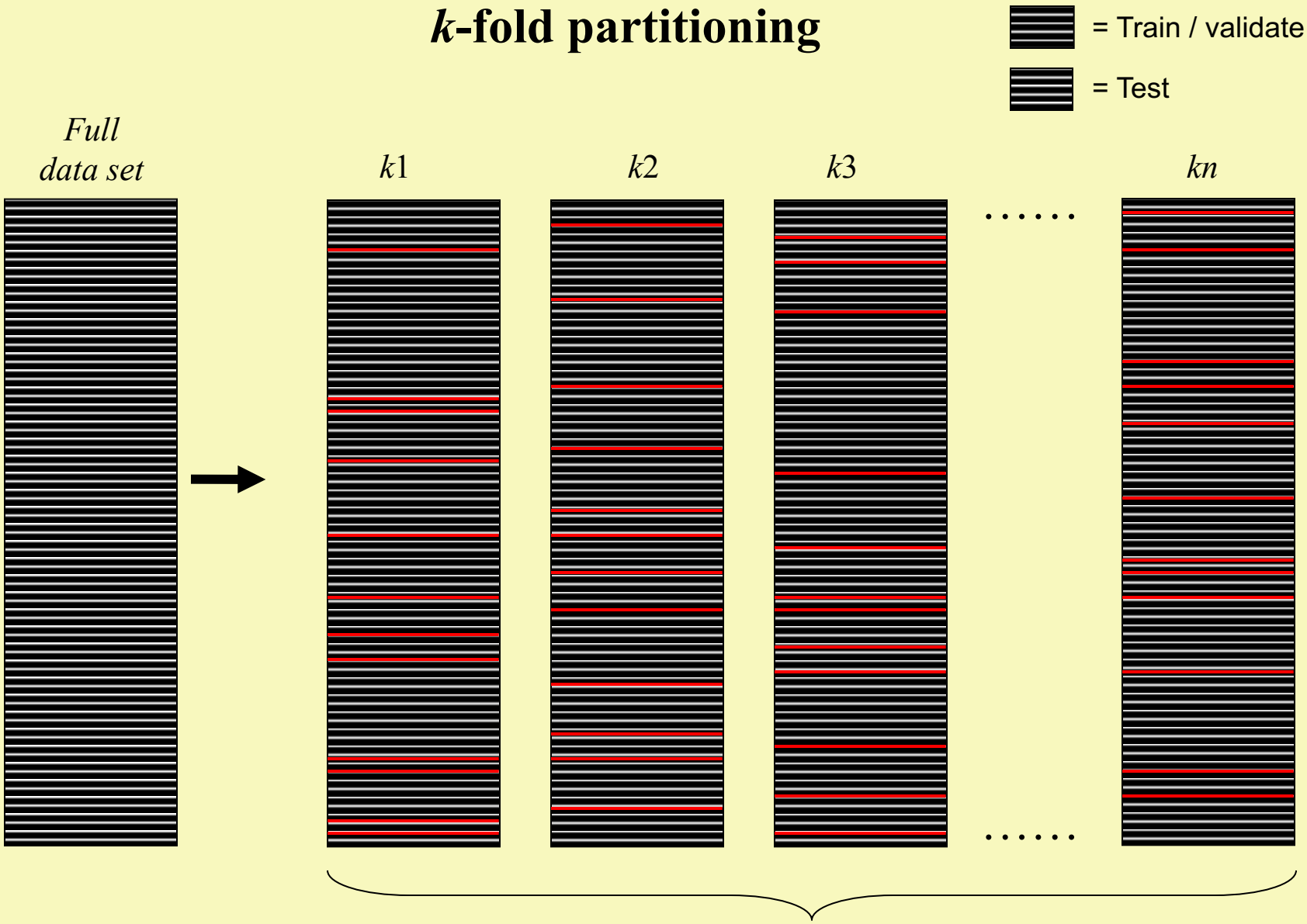
Cross validation

1. Split data randomly into k roughly equal-sized parts. Take turns using each part as a test set and the other $k - 1$ parts for model training.
2. Compute test statistic each time. Cross-validation estimate of predictive performance is the average of the k tests.



(C. Graham)

Model calibration and evaluation strategies: *k*-fold partitioning

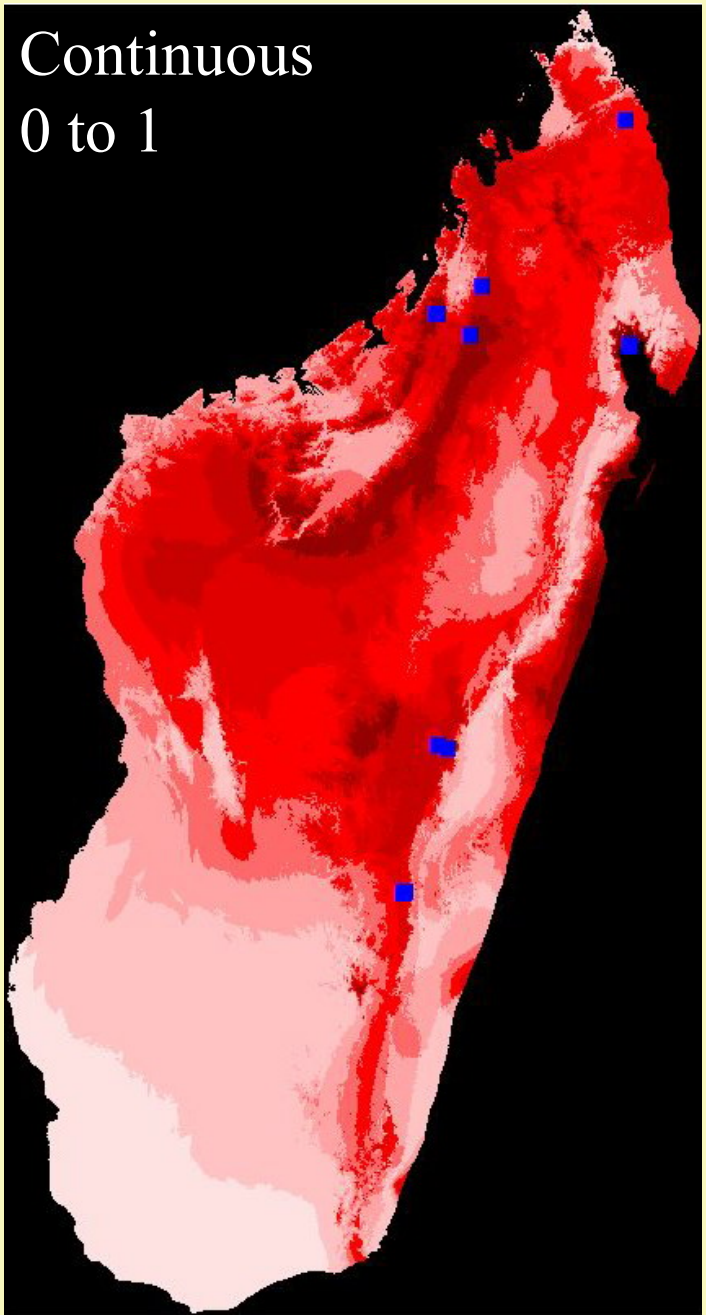


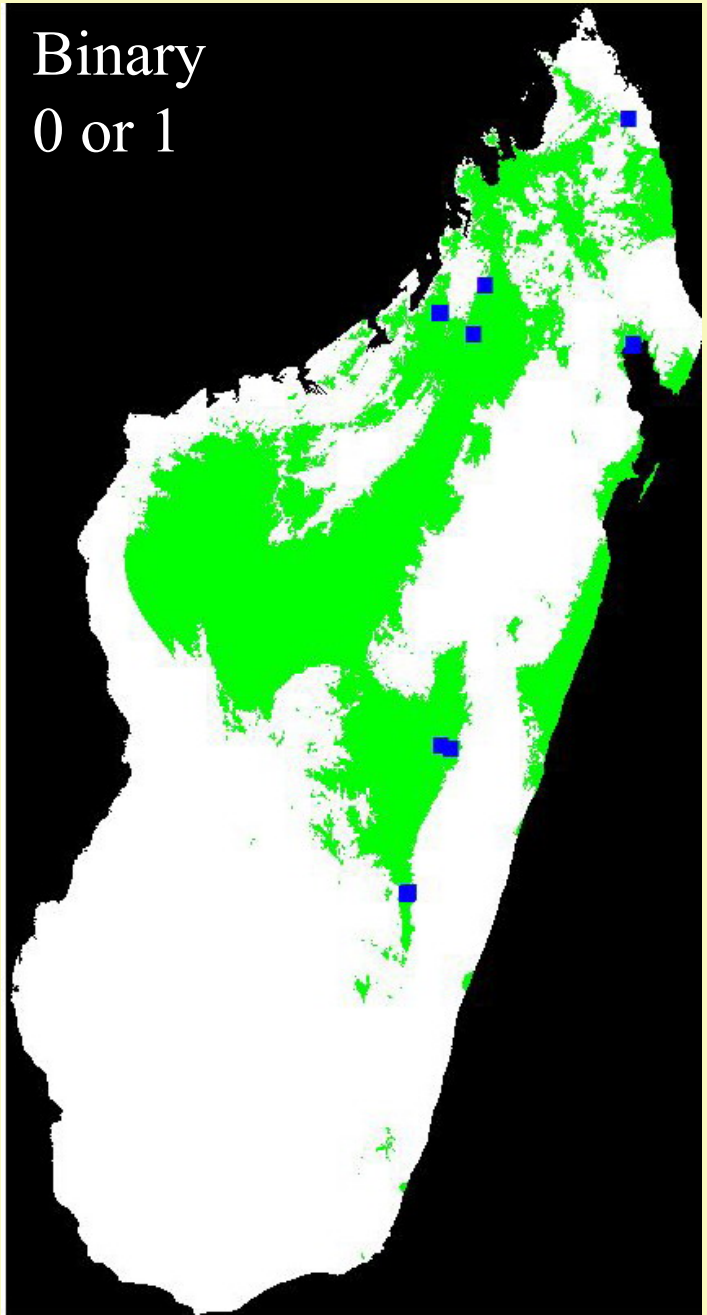
2 main types of test:

- Threshold dependent
- Threshold independent

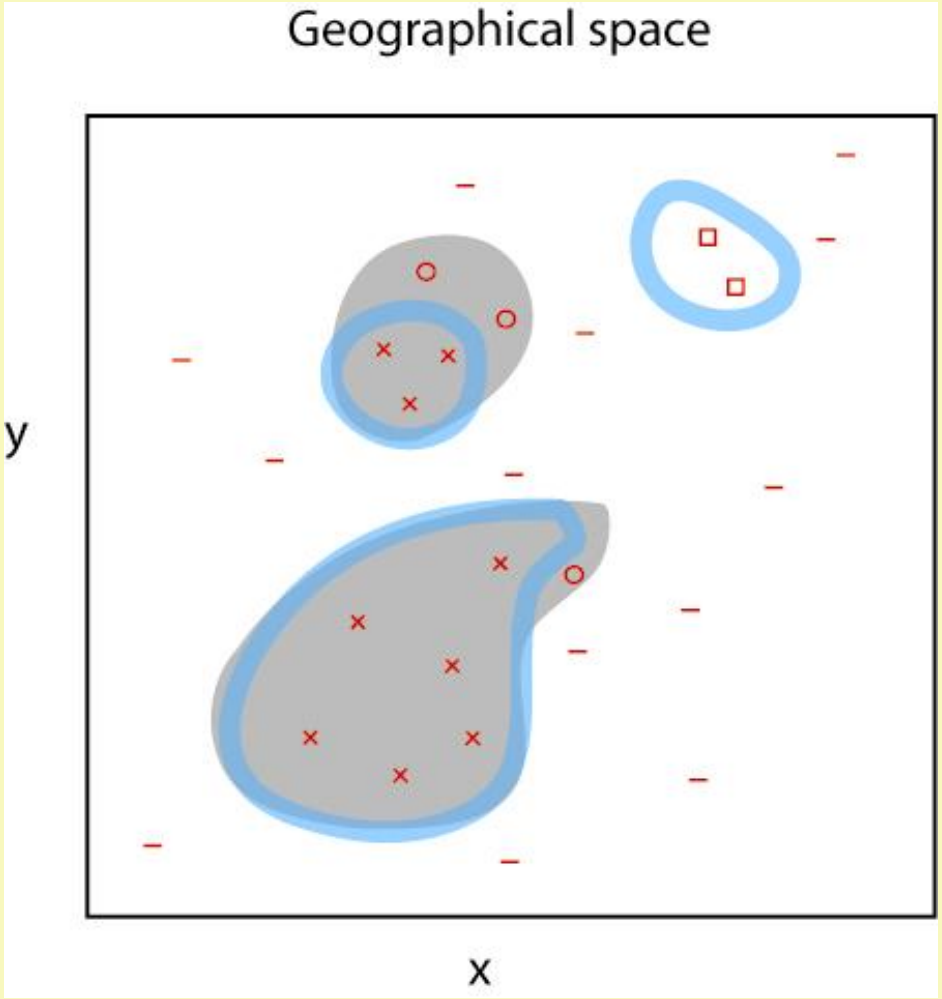
Used to tell which part of a model is useful & whether one model is better than another.

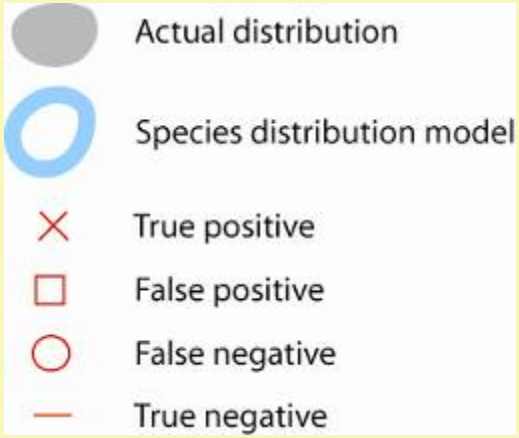
THRESHOLDING





The four types of results that are possible when testing a distribution model





(see Pearson NCEP module 2007)

Presence-absence confusion matrix

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Presence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed presences correctly predicted
(or '**sensitivity**', or 'true positive fraction'):

$$a/(a + c)$$

Presence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed presences correctly predicted
(or '**sensitivity**', or 'true positive fraction'):

$$a/(a + c)$$

Proportion of observed presences incorrectly predicted
(or '**omission** rate', or 'false negative fraction'):

$$c/(a + c)$$

Absence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed (or assumed) absences correctly predicted
(or '**specificity**', or 'true negative fraction'):

$$d/(b + d)$$

Absence-only test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion of observed (or assumed) absences correctly predicted (or '**specificity**', or 'true negative fraction'):

$$d/(b + d)$$

Proportion of observed (or assumed) absences incorrectly predicted (or '**commission** rate', or 'false positive fraction'):

$$b/(b + d)$$

Presence-absence test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion correctly predicted, or ‘accuracy’, or ‘correct classification rate’:

$$(a + d)$$

Presence-absence test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Proportion correctly predicted, or ‘accuracy’, or ‘correct classification rate’:

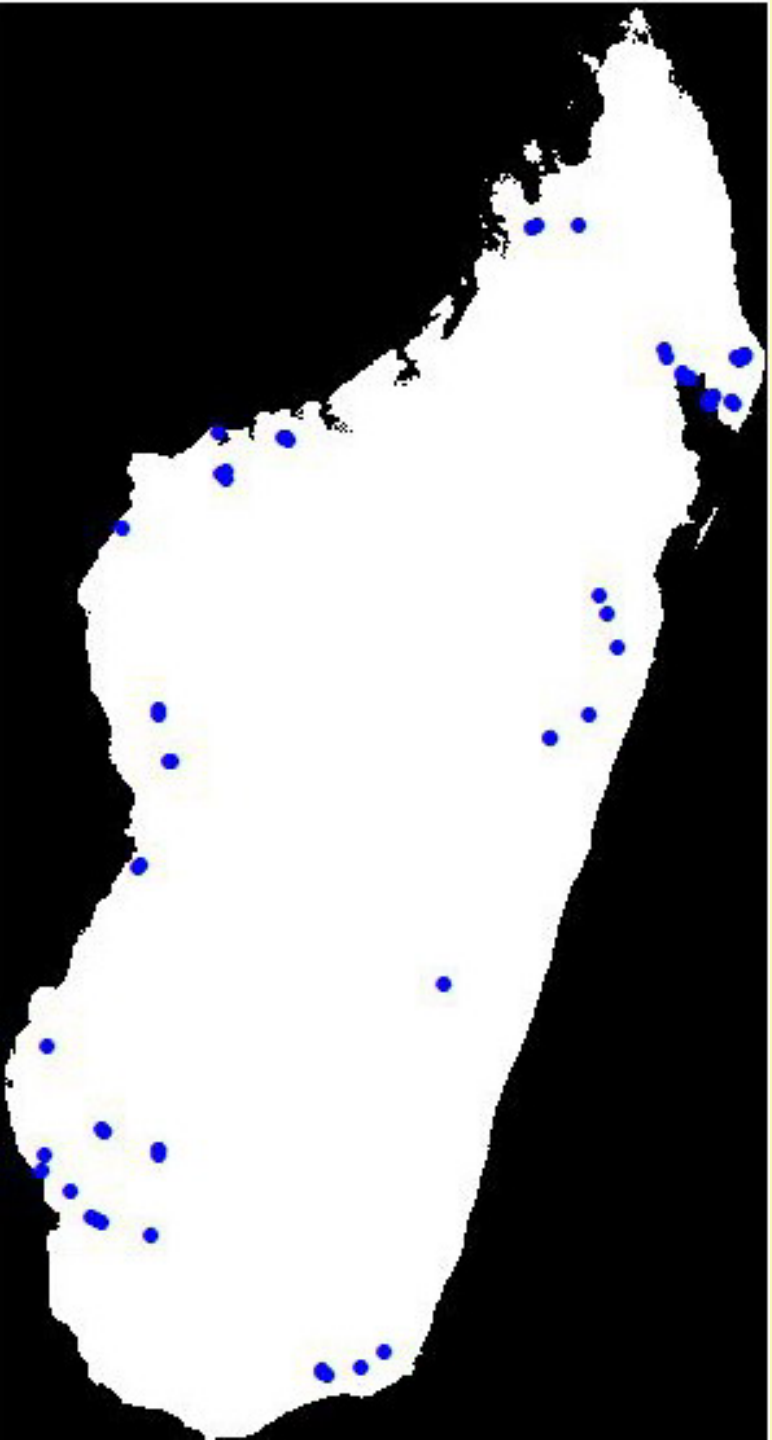
$$(a + d)/(a + b + c + d)$$

WHICH THRESHOLD IS BEST?

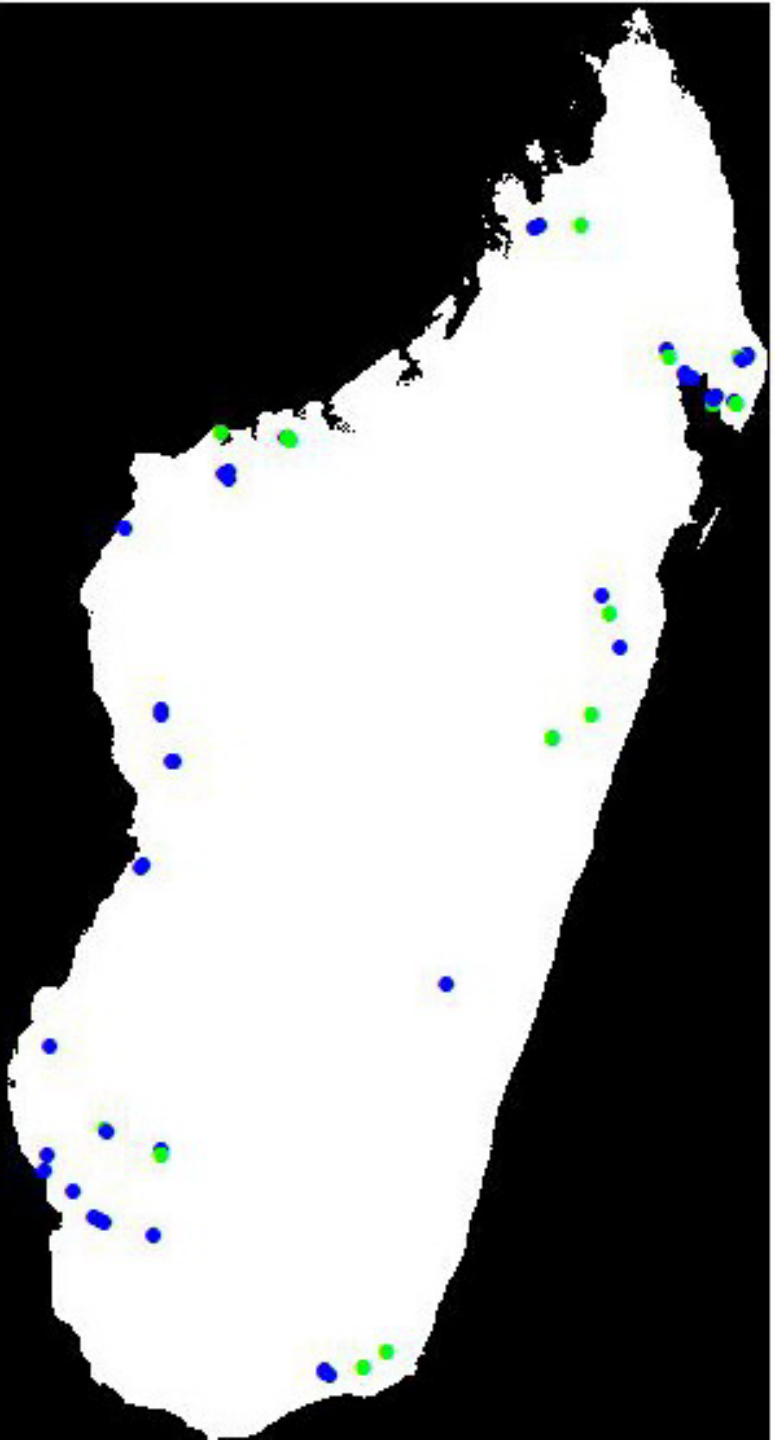
A very complicated answer!

Again, it depends on what you want to use the model for.

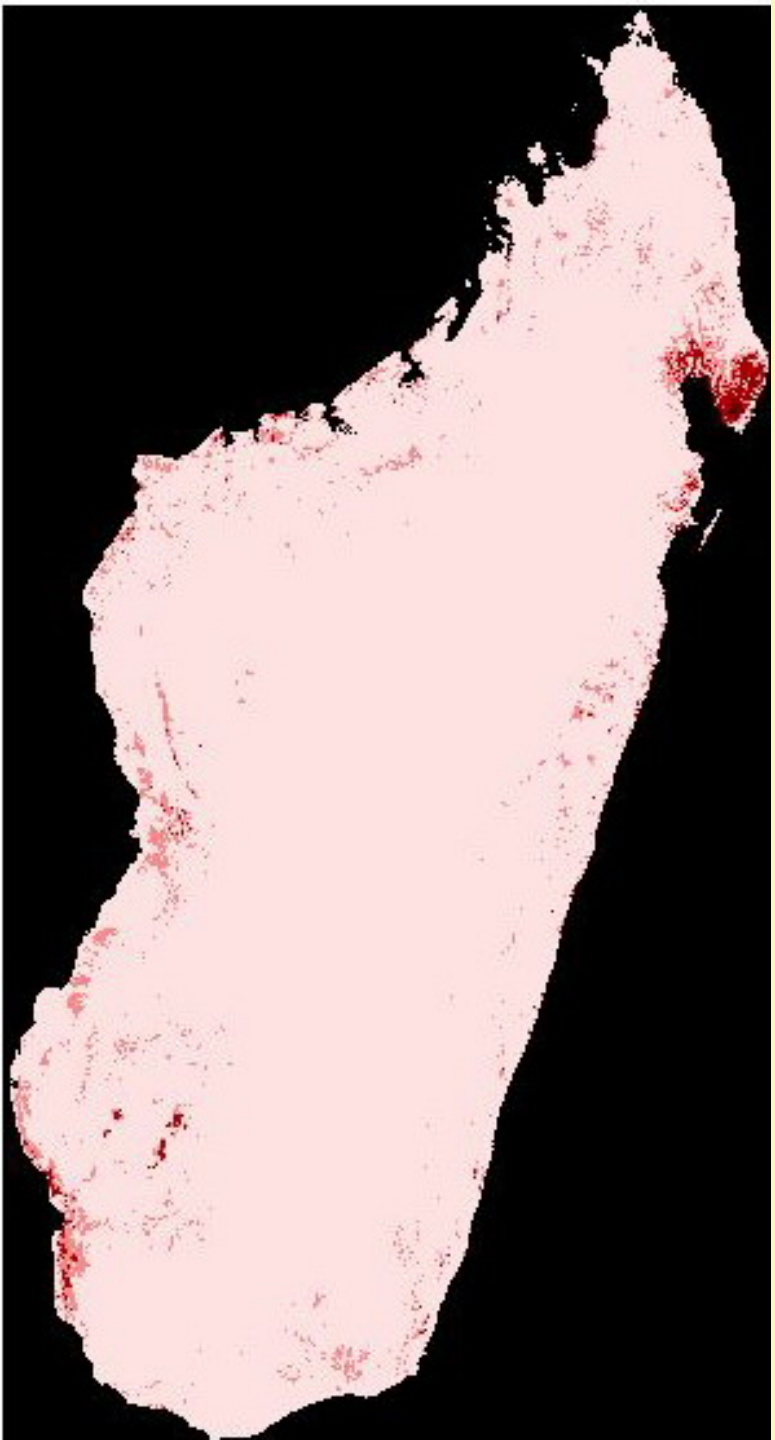
TO COMPARE THRSHOLDS:



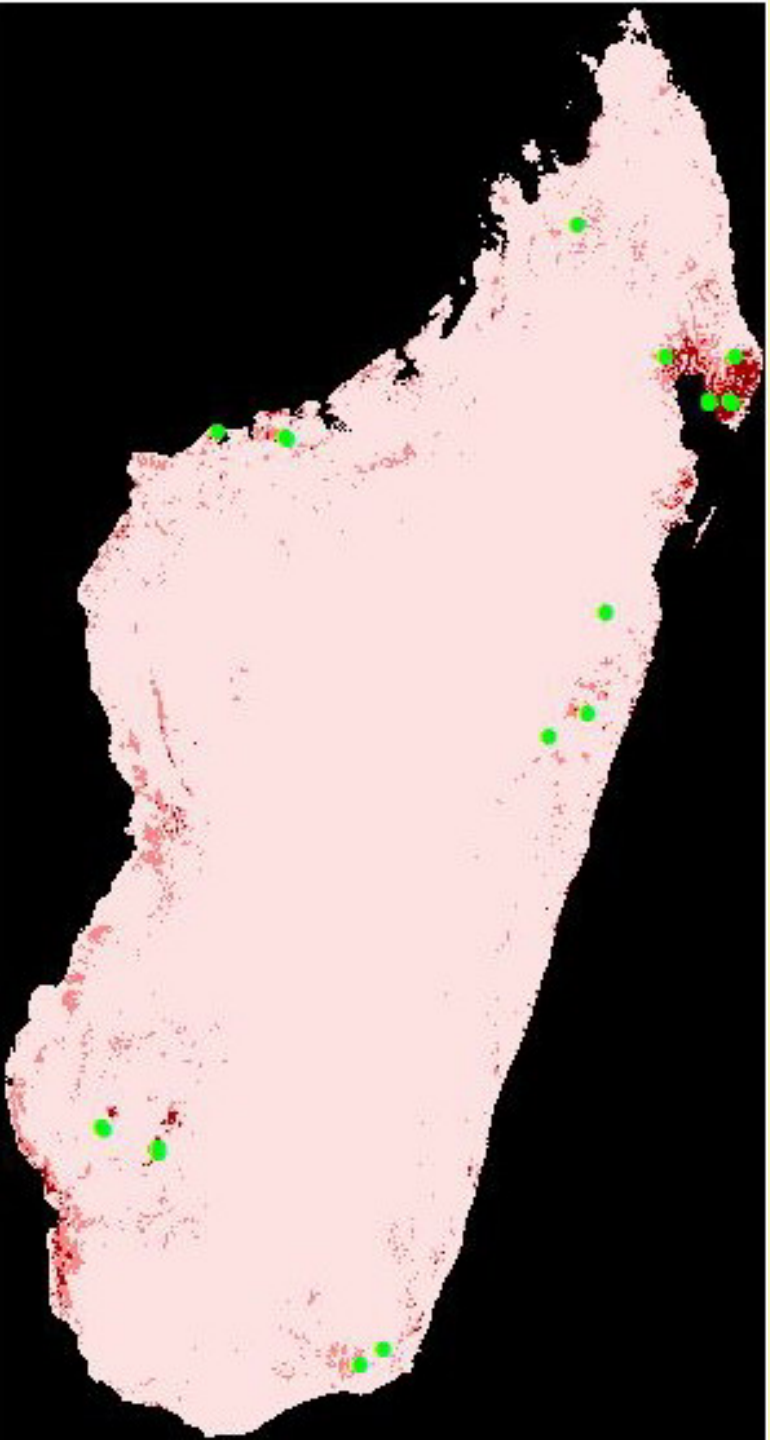
All Presence Data



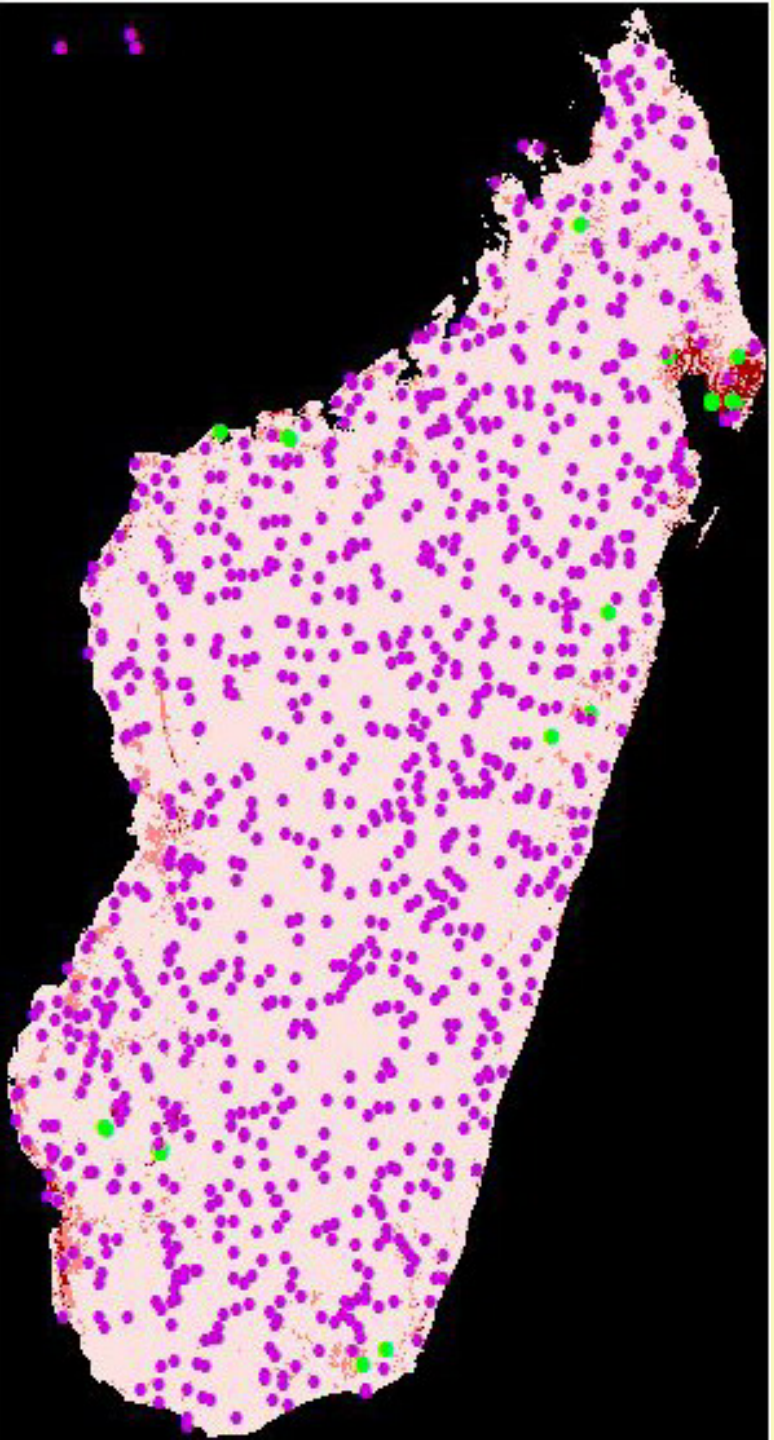
Split presence data:
75% Training
25% Testing



Derive Model from
75% Training data

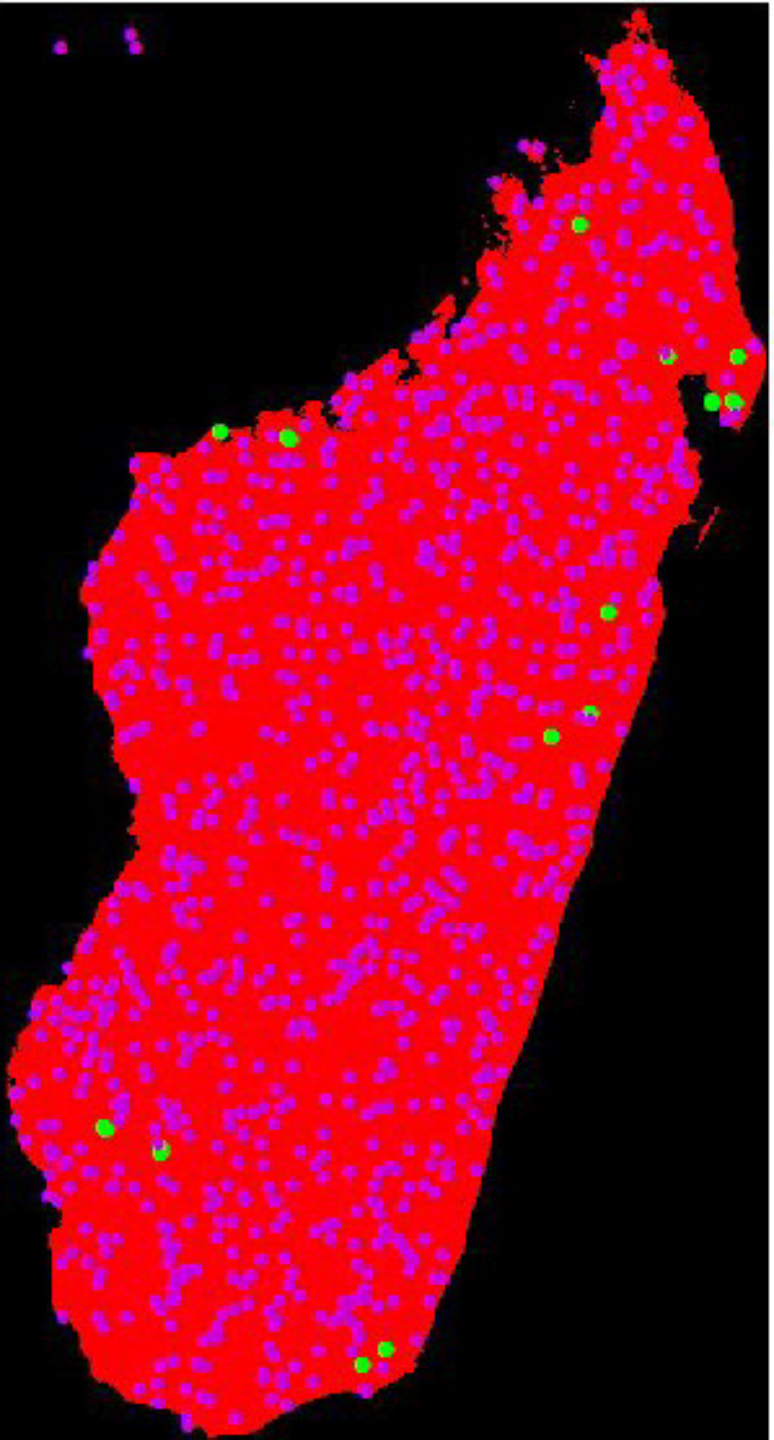


Overlay Presence Test Data



Overlay Absence Test Data

- in our case we use randomly generated “pseudo-absences”

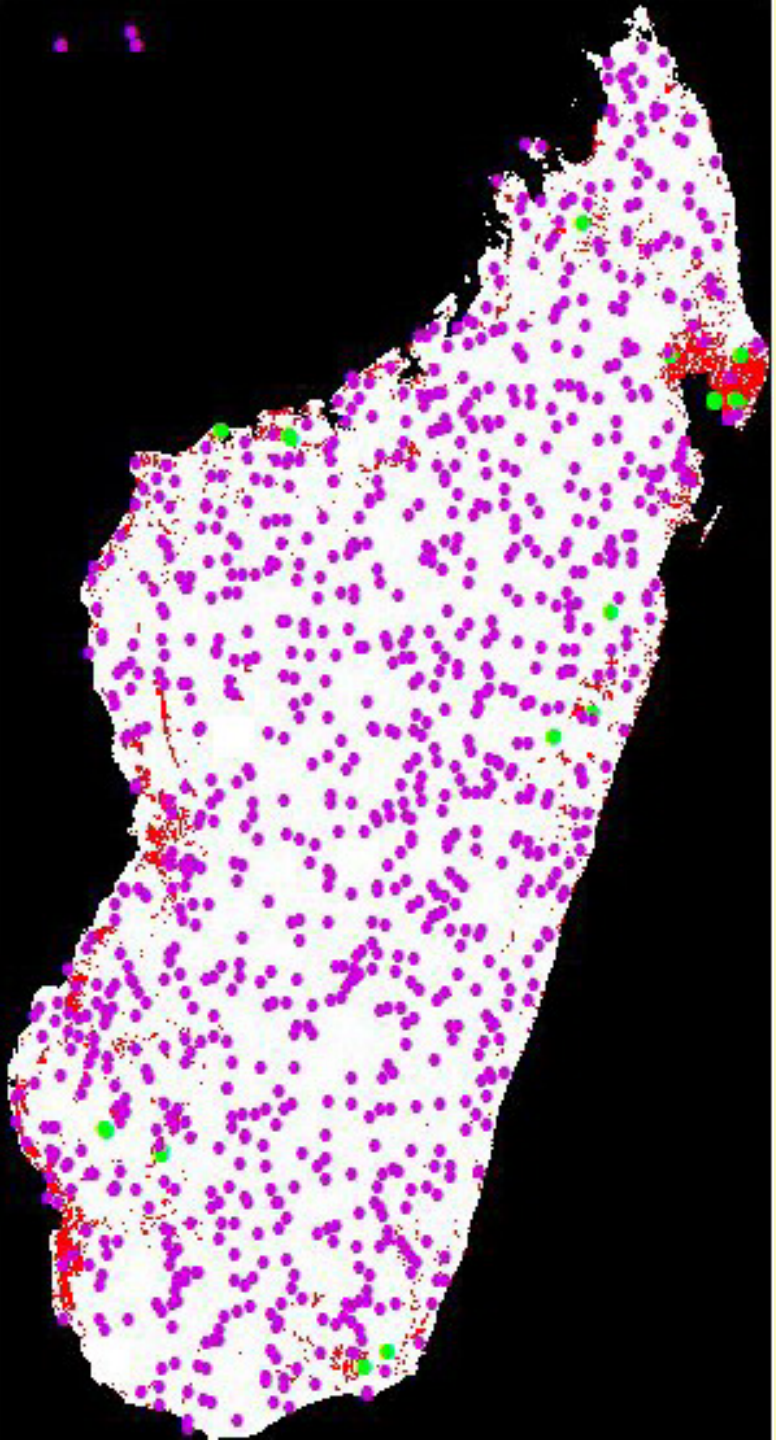


First Threshold 0.0

		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives

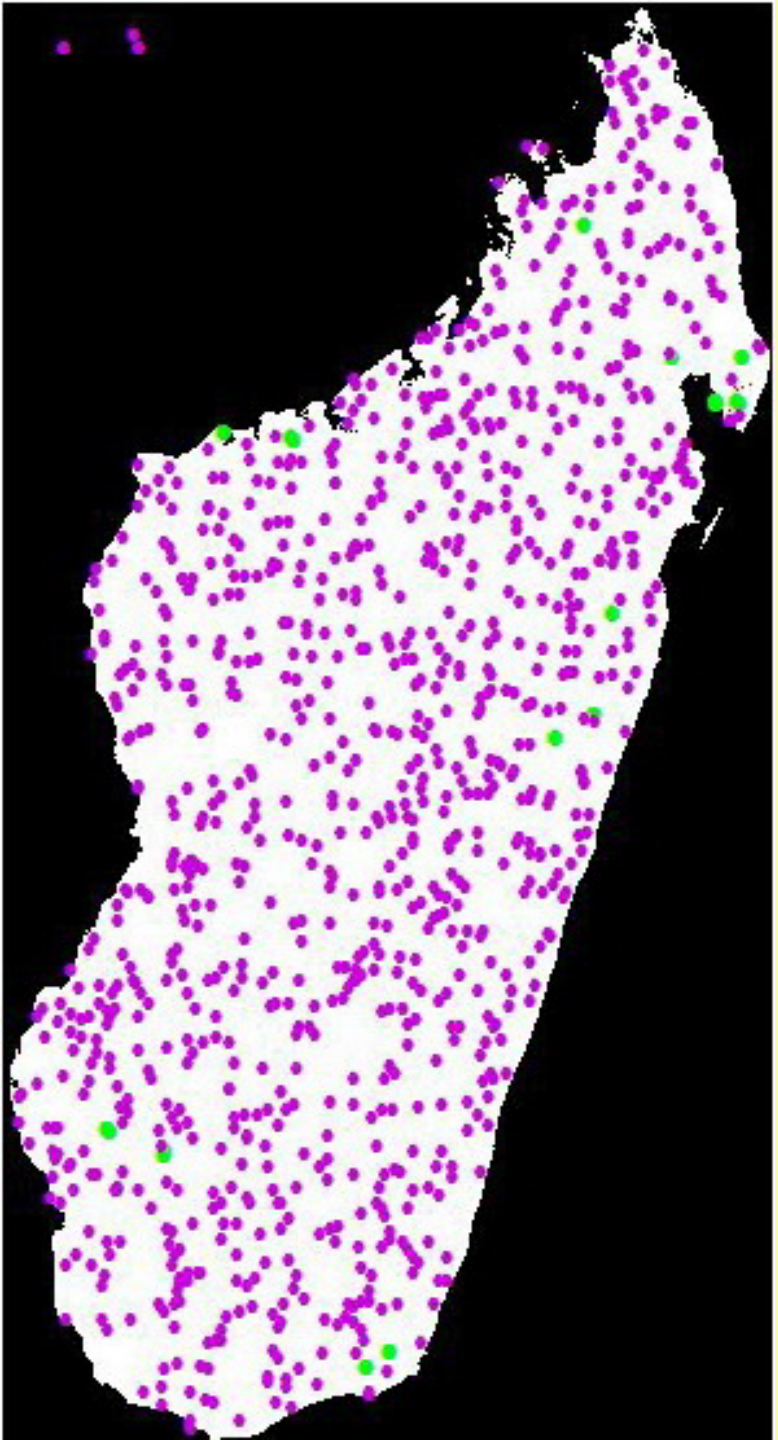


Second Threshold 0.75

		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives



Third Threshold 0.9

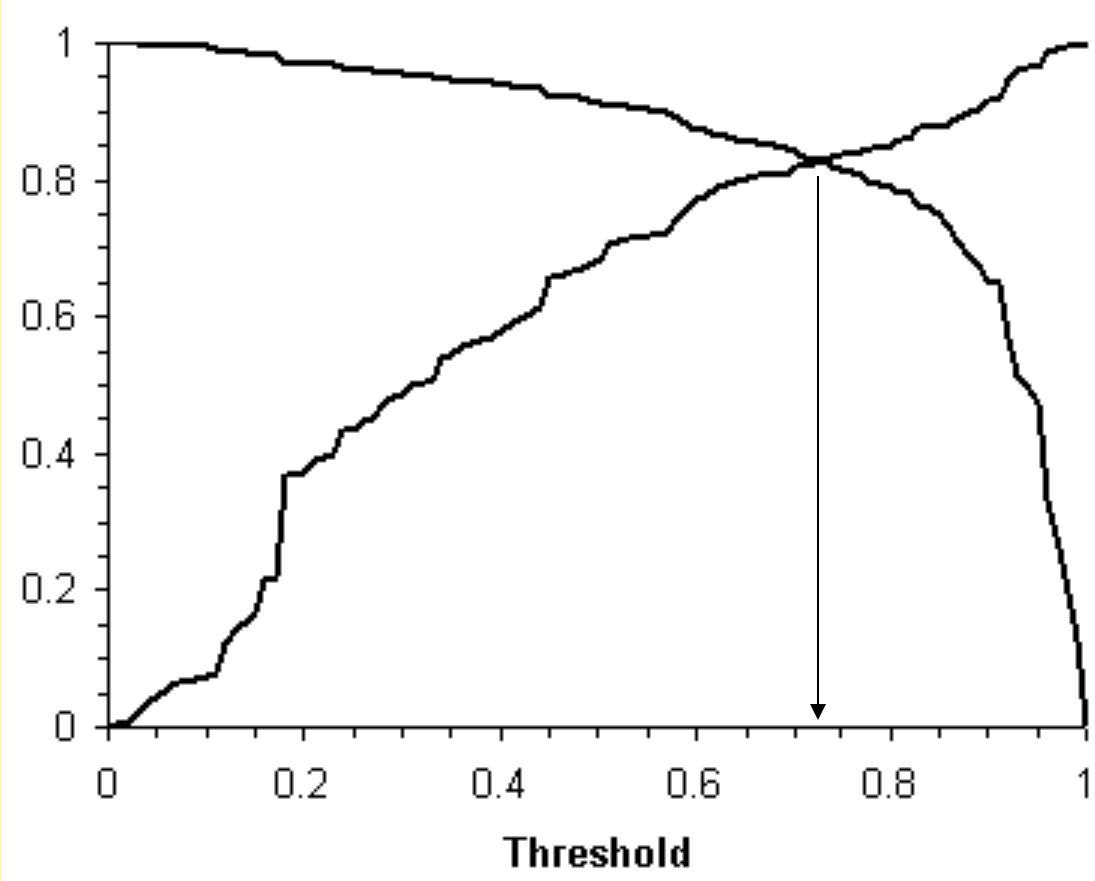
		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives

WHICH THRESHOLD IS BEST?

presence-absence data



Omission (c/a+c)
(false negative fraction
i.e. proportion of
presences predicted
absent)

Commission (b/b+d)
(false positive fraction
i.e. proportion of
absences predicted
present)

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

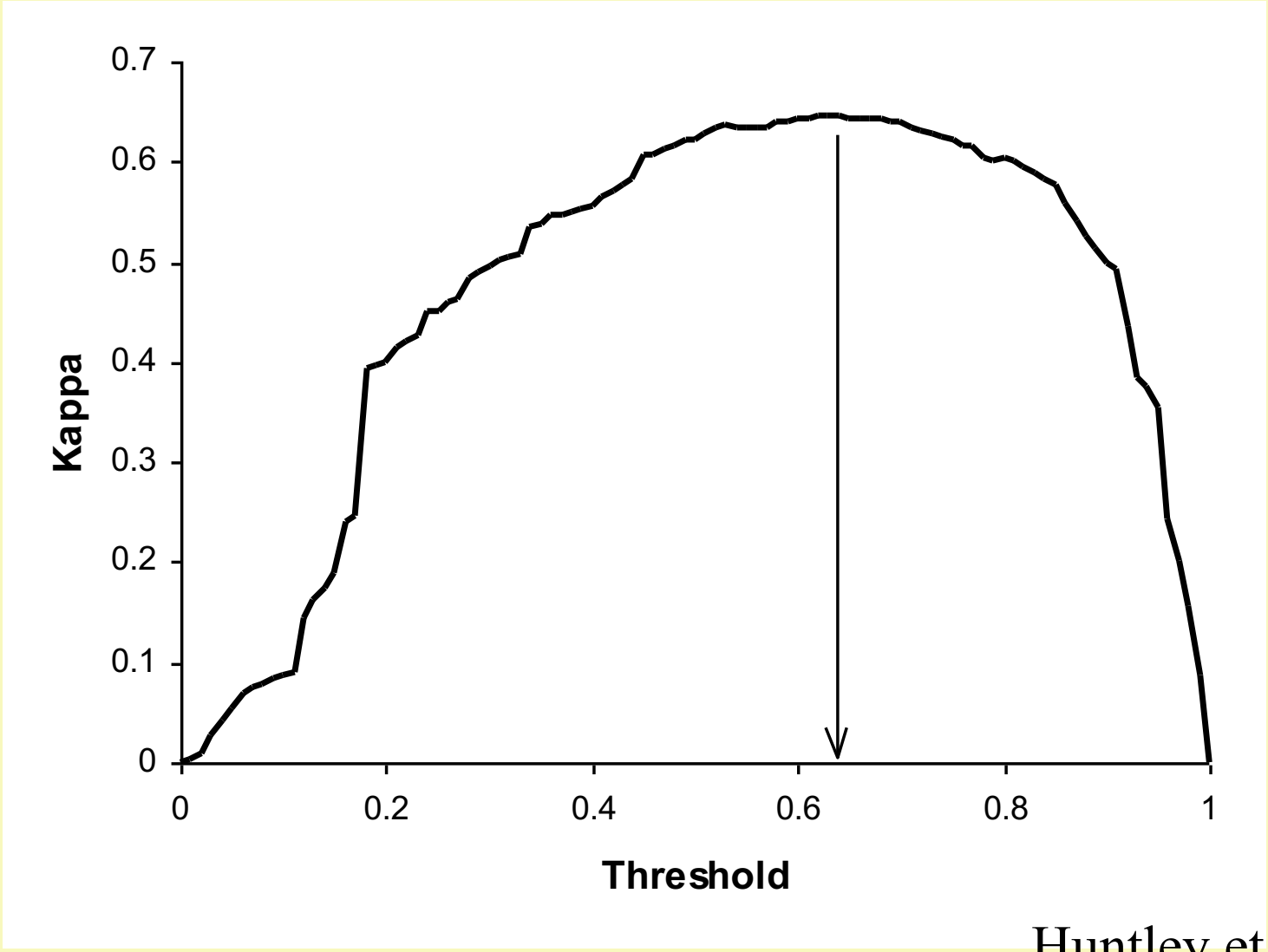
Presence-absence test statistics

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Cohen's Kappa:

$$k = \frac{[(a+d) - (((a+c)(a+b)+(b+d)(c+d))/n)]}{[n - (((a+c)(a+b)+(b+d)(c+d))/n)]}$$

Selecting a decision threshold (p/a data)



Huntley et al. 1995

Threshold selection

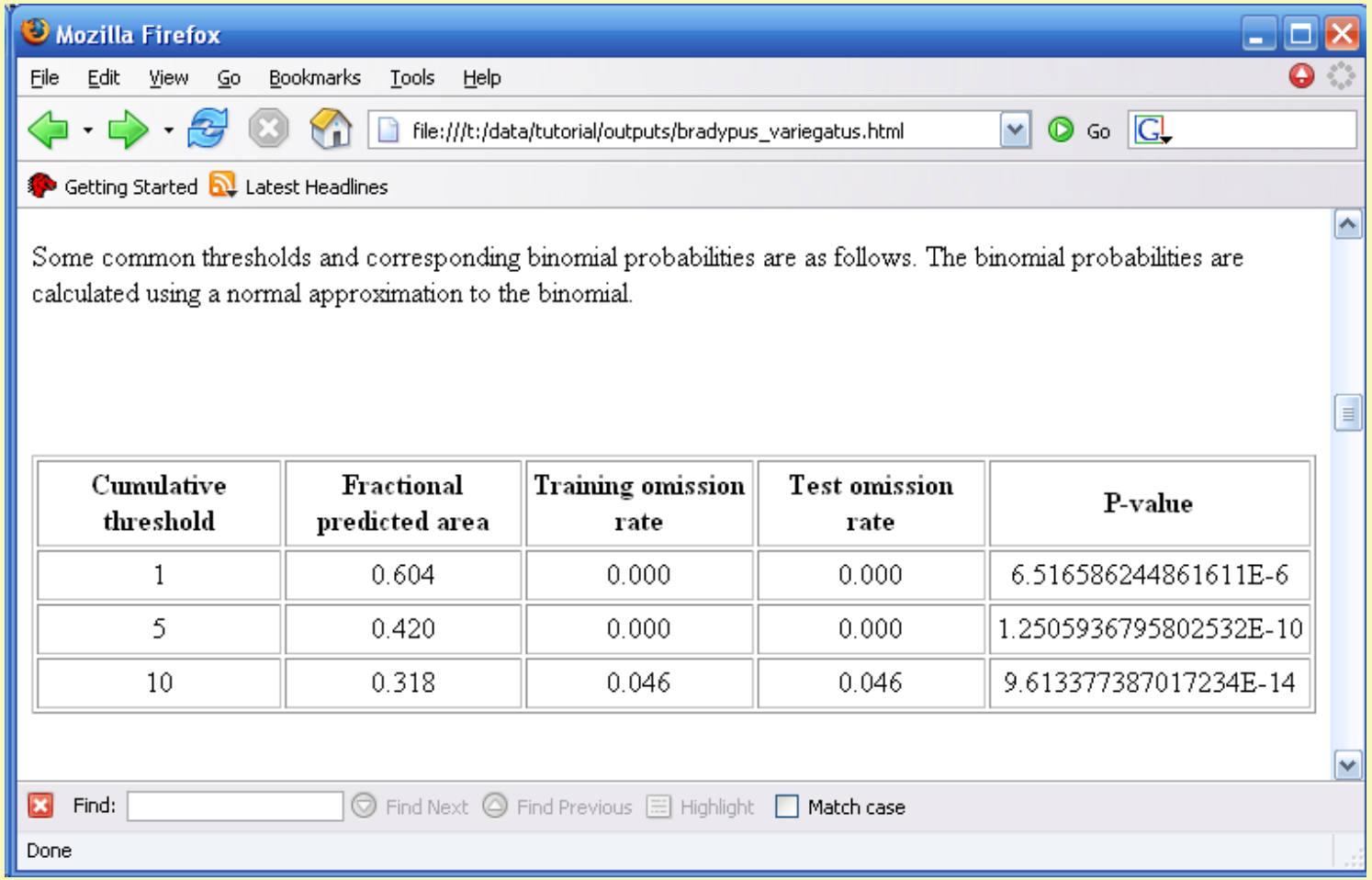
- Liu, C., Berry, P. M., Dawson, T. P. and Pearson, R. G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385-/393.

Selecting a decision threshold (p/a data)

Code	Approach	Definition	Reference
Subjective approach			
1	Fixed threshold approach	Taking a fixed value, usually 0.5, as the threshold	Manel et al. (1999), Bailey et al. (2002)
Objective approaches			
	Single index-based approaches:		
2	Kappa maximization approach	Kappa statistic is maximized	Huntley et al. (1995), Guisan et al. (1998)
3	OPS maximization approach	Overall prediction success (OPS) is maximized	
	Model-building data-only-based approach:		
4	Prevalence approach	Taking the prevalence of model-building data as the threshold	Cramer (2003)
	Predicted probability/suitability-based approaches:		
5	Average probability/suitability approach	Taking the average predicted probability/ suitability of the model-building data as the threshold	Cramer (2003)
6	Mid-point probability/suitability approach	Mid-point between the average probabilities of or suitabilities for the species' presence for occupied and unoccupied sites	Fielding and Haworth (1995)
	Sensitivity and specificity-combined approaches:		
7	Sensitivity-specificity sum maximization approach	The sum of sensitivity and specificity is maximized	Cantor et al. (1999), Manel et al. (2001)
8	Sensitivity-specificity equality approach	The absolute value of the difference between sensitivity and specificity is minimized	Cantor et al. (1999)
9	ROC plot-based approach	The threshold corresponds to the point on ROC curve (sensitivity against 1-specificity) which has the shortest distance to the top-left corner (0,1) in ROC plot	Cantor et al. (1999)
	Precision and recall-combined approaches:		
10	Precision-recall break-even point approach	The absolute value of the difference between precision and recall is minimized	Shapire et al. (1998)
11	P-R plot-based approach	The threshold corresponds to the point on P-R (Precision-Recall) curve which has the shortest distance to the top-right corner (1,1) in P-R plot	
12	F maximization approach	The index F is maximized. In this study, $\alpha = 0.5$ is used in F, i.e. there is no preference to precision and recall	Shapire et al. (1998)

(Liu et al. 2005 *Ecography* 29:385-393)

Maxent Output



Cumulative threshold	Fractional predicted area	Training omission rate	Test omission rate	P-value
1	0.604	0.000	0.000	6.516586244861611E-6
5	0.420	0.000	0.000	1.2505936795802532E-10
10	0.318	0.046	0.046	9.613377387017234E-14

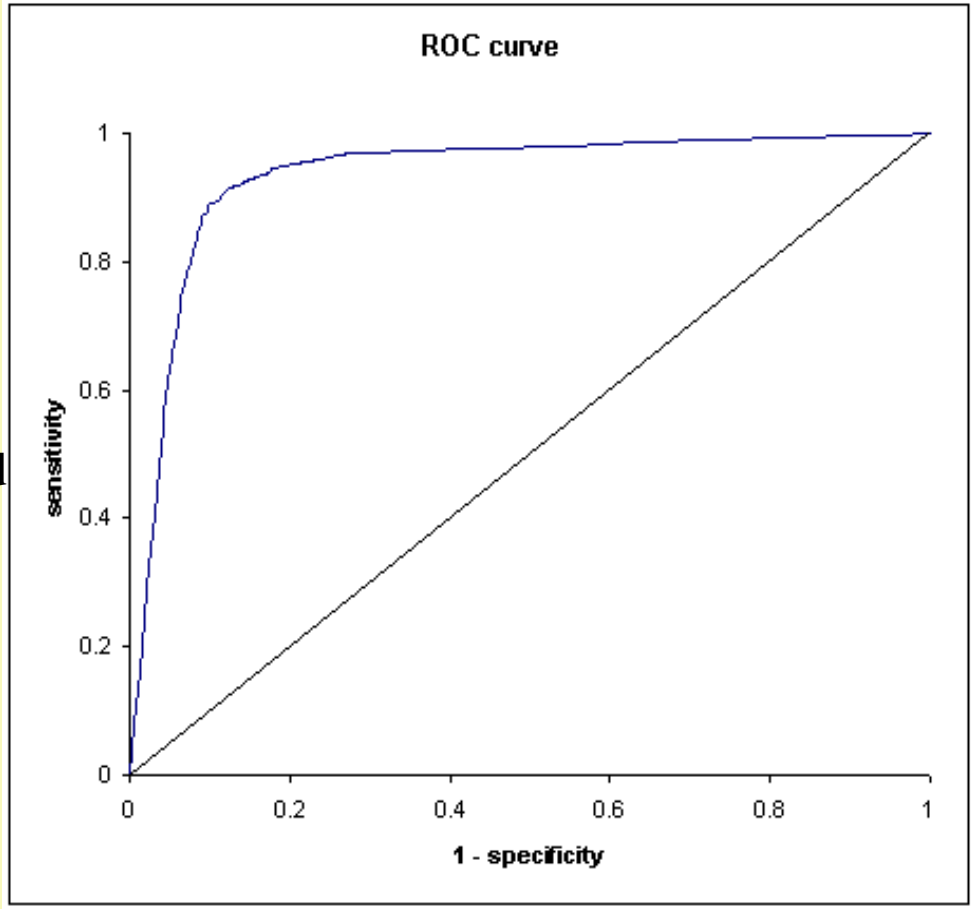
Area Under the Reciever Operator Characteristic Curve (AUC)

A threshold-independent test statistic:

A Reciever Operator Characteristic (ROC) plot

sensitivity = a/(a+c)

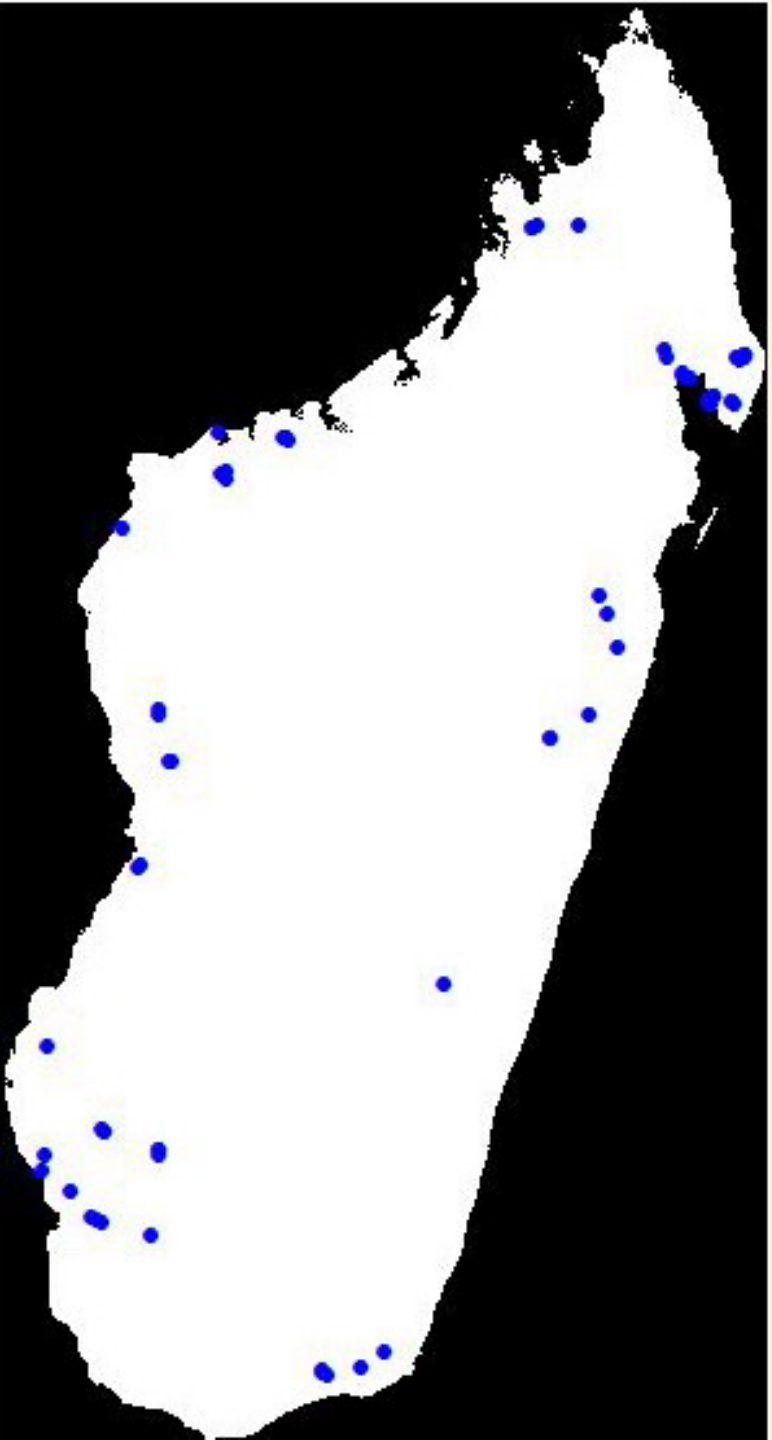
Fraction of presences predicted present.



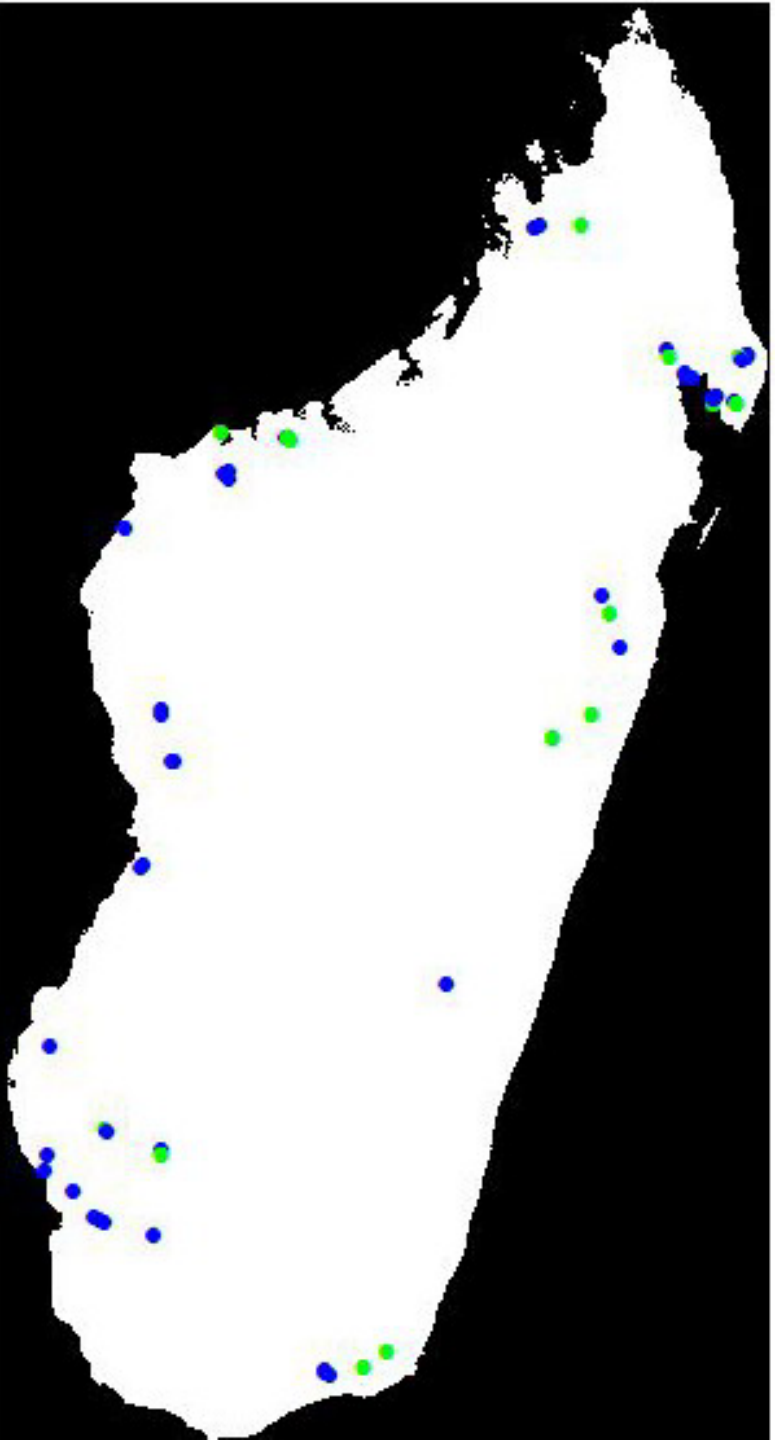
1- specificity = 1 – [d/(b+d)]

Fraction of absences predicted present

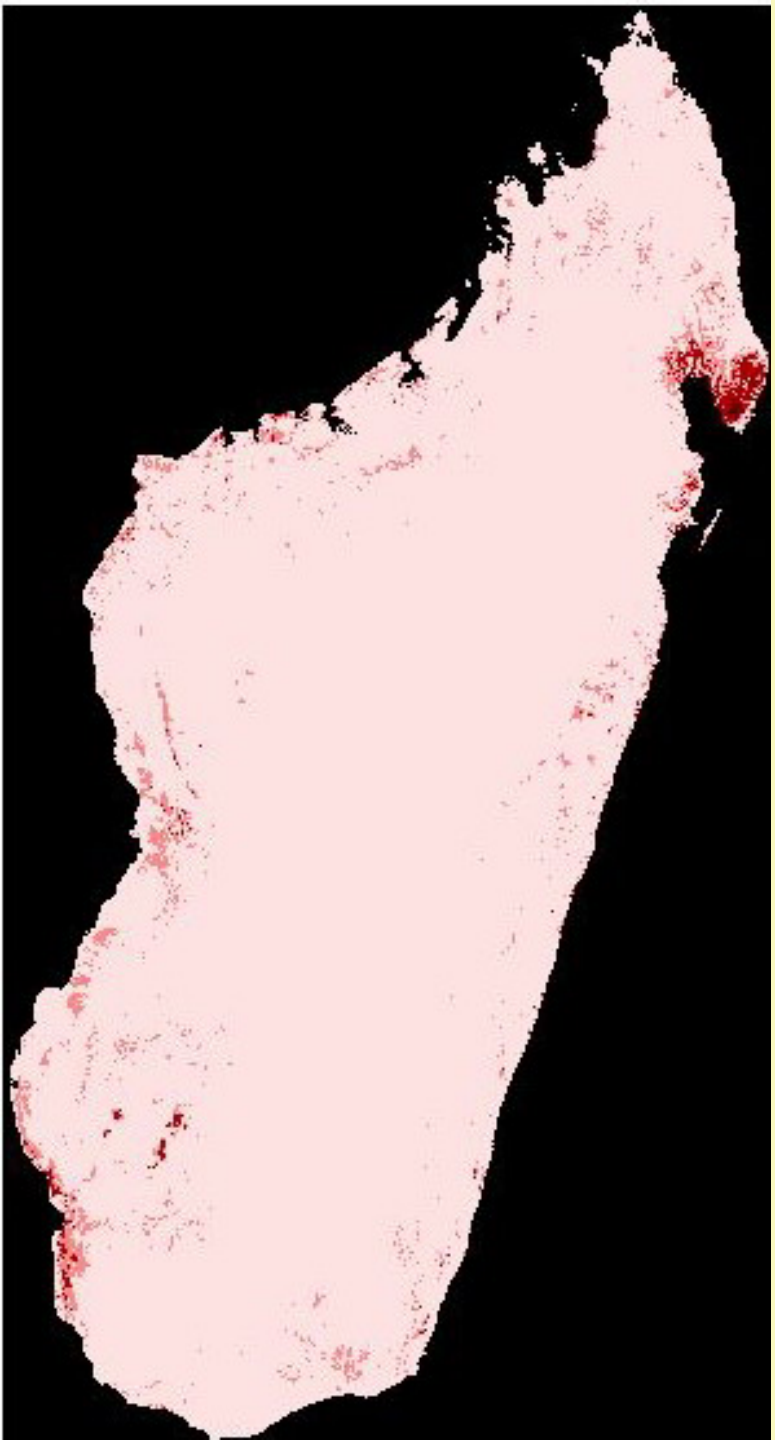
	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)



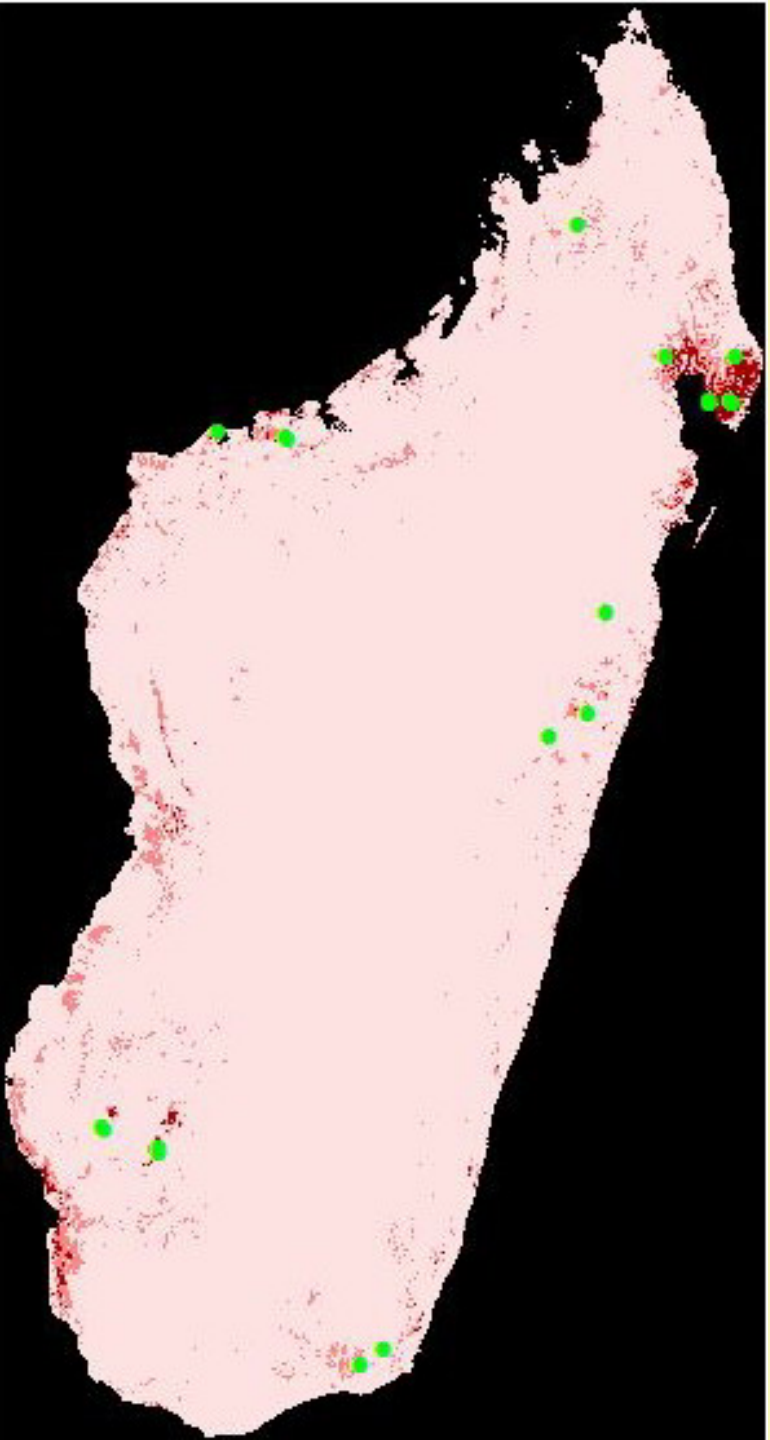
All Presence Data



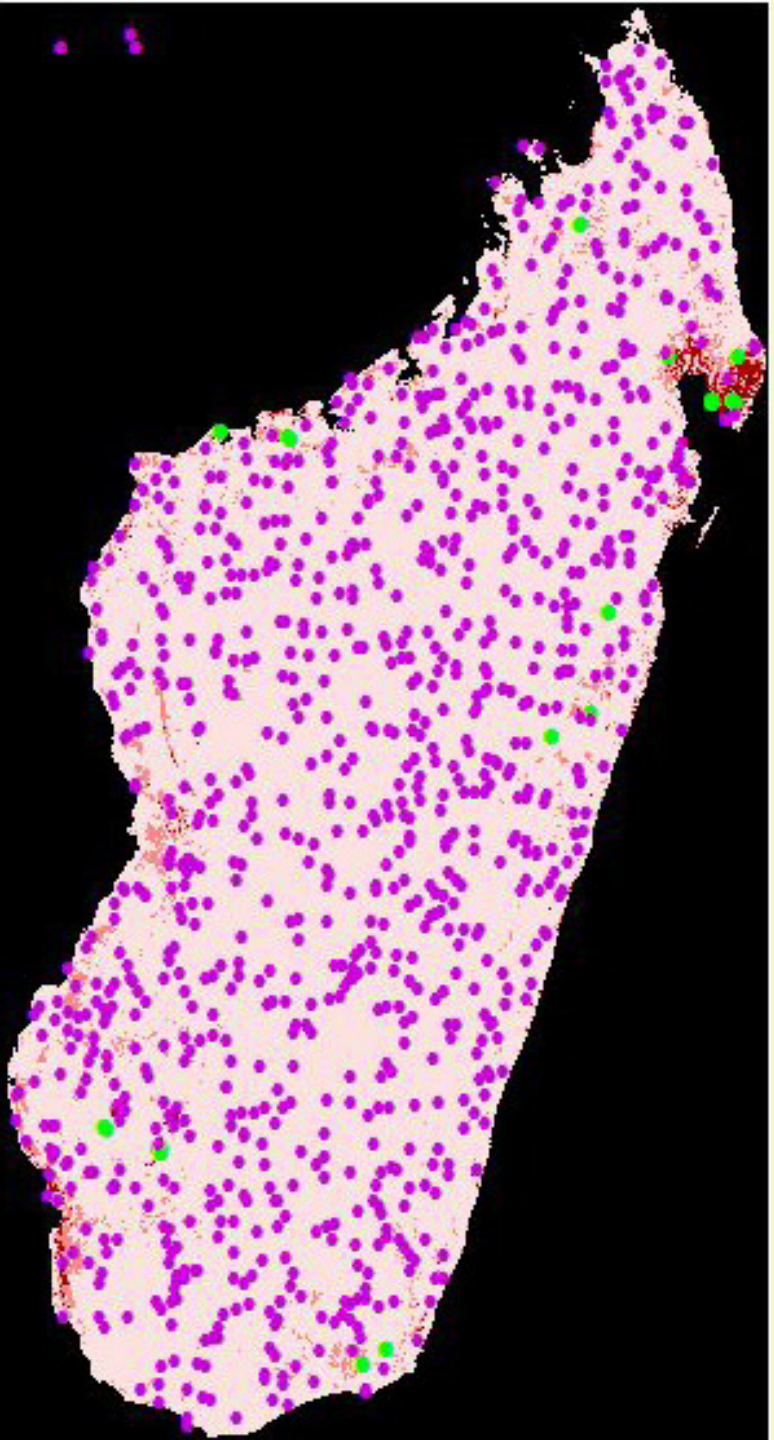
Split presence data:
75% Training
25% Testing



Derive Model from
75% Training data

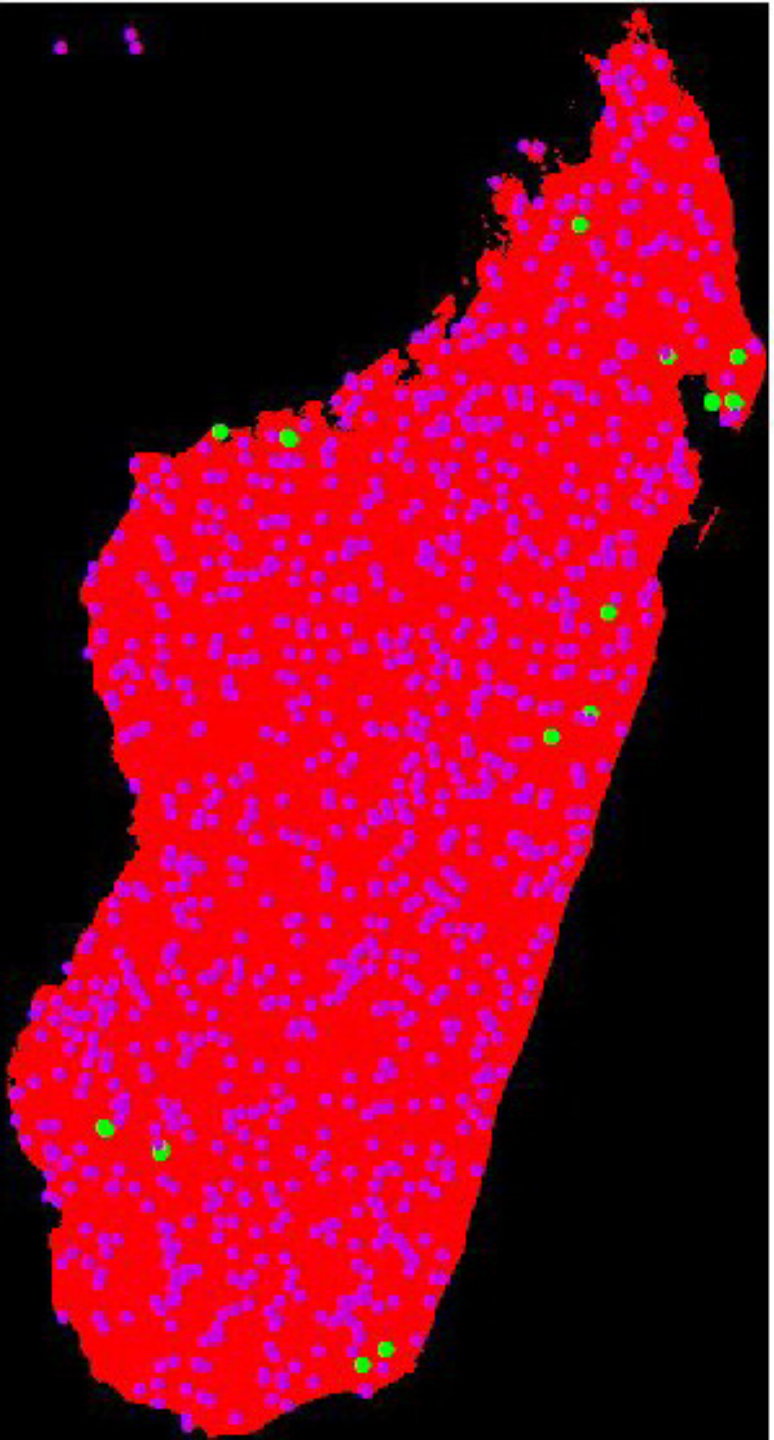


Overlay Presence Test Data



Overlay Absence Test Data

- in our case we use randomly generated “pseudo-absences”



First Threshold 0.0

		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

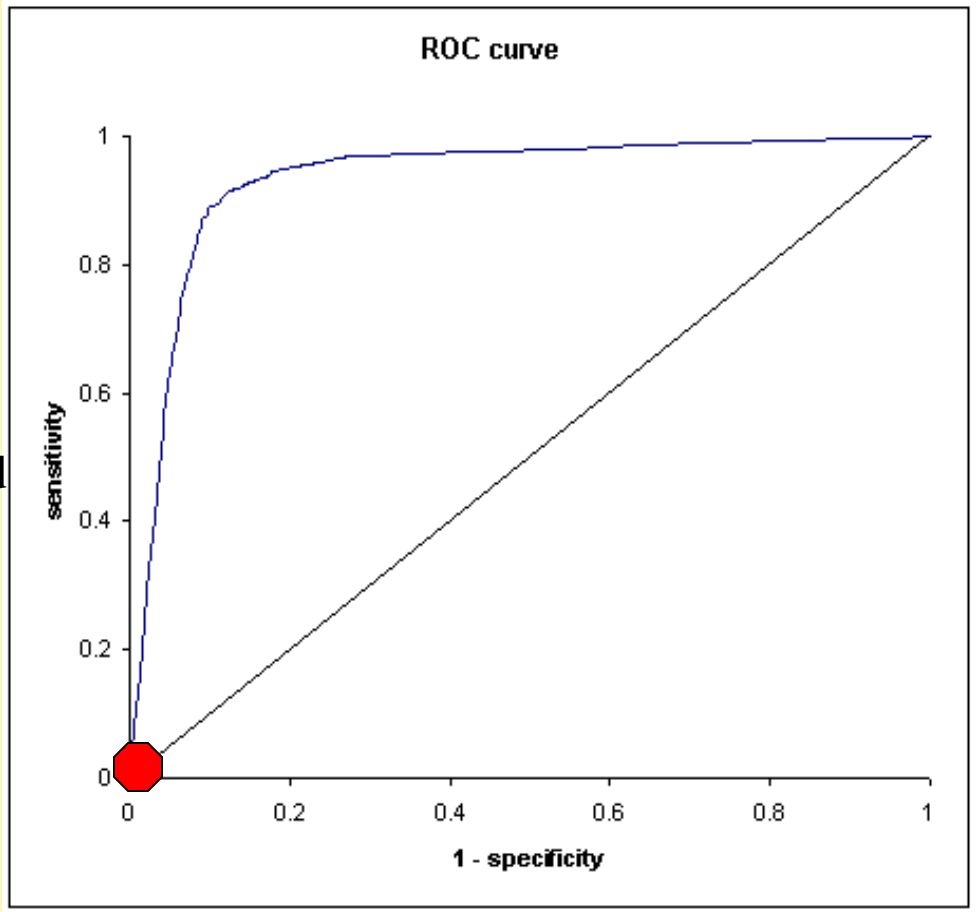
Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives

A Reciever Operator Characteristic (ROC) plot

sensitivity = a/(a+c)

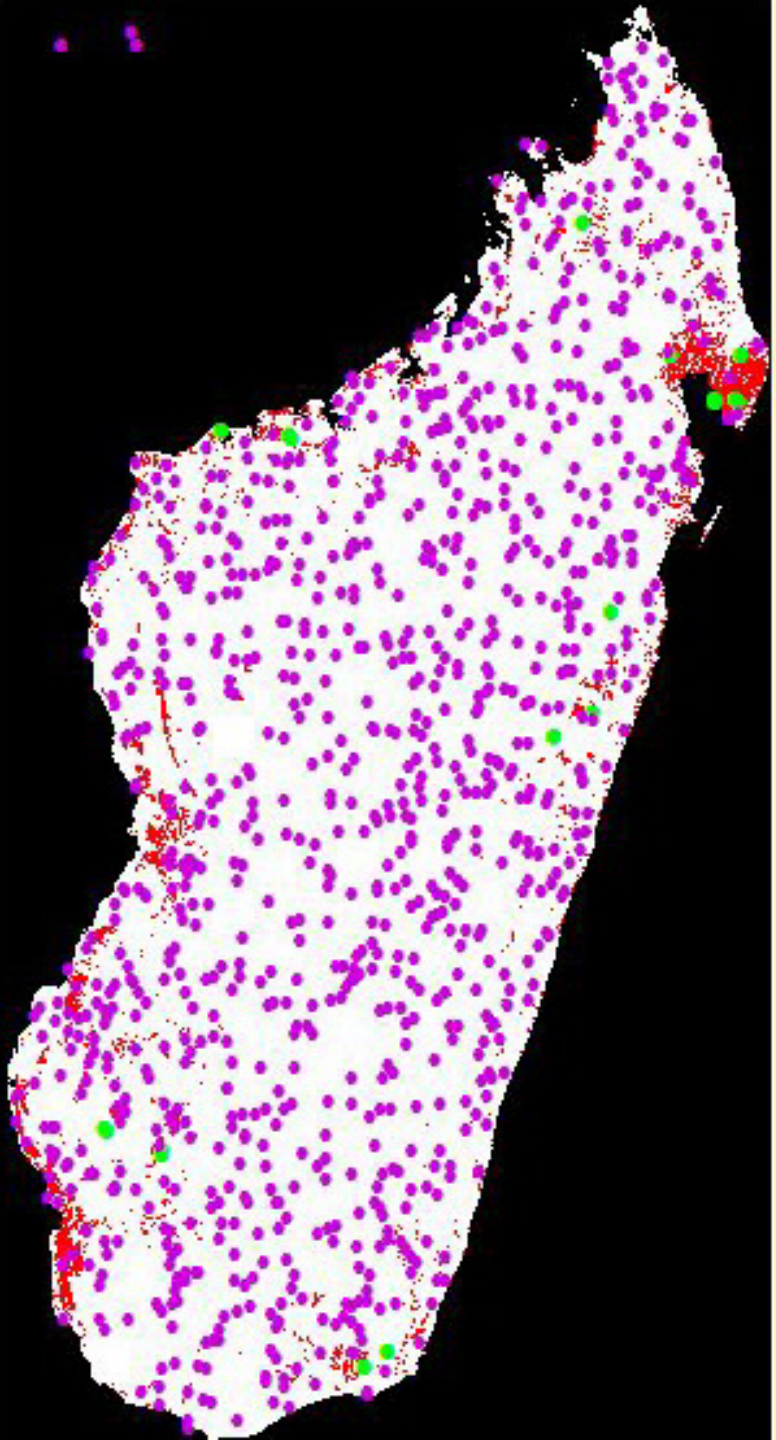
Fraction of presences predicted present.



1- specificity = 1 – [d/(b+d)]

Fraction of absences predicted present

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)



Second Threshold 0.75

		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

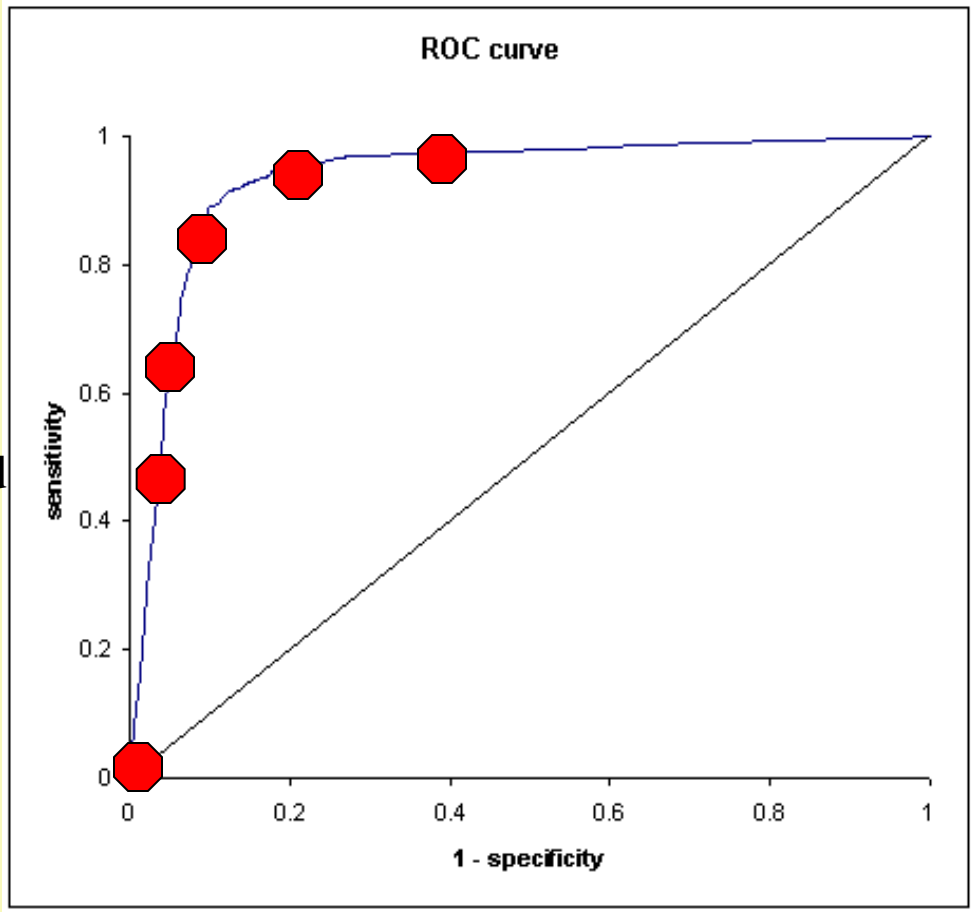
Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives

A Reciever Operator Characteristic (ROC) plot

sensitivity = a/(a+c)

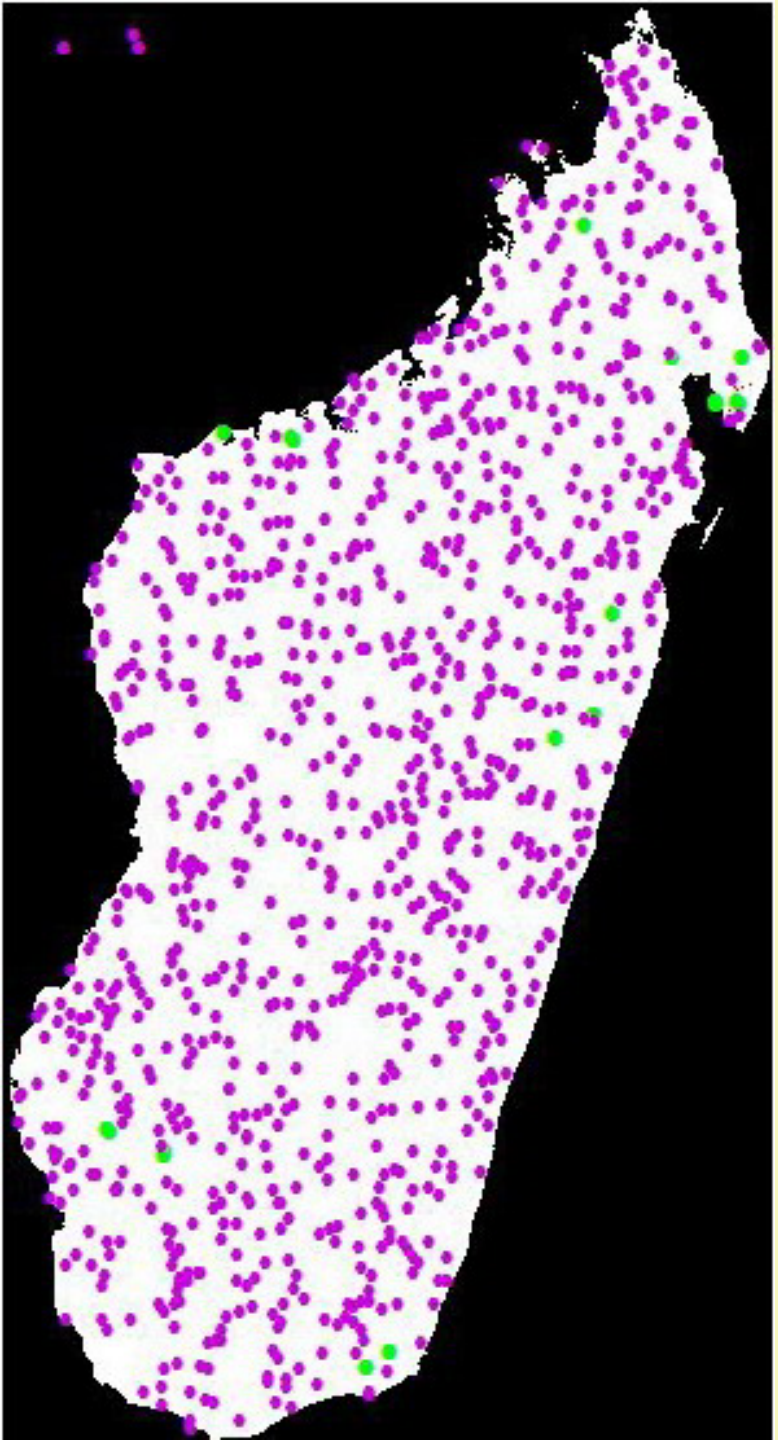
Fraction of presences predicted present.



1- specificity = 1 – [d/(b+d)]

Fraction of absences predicted present

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)



Third Threshold 0.9

		ACTUAL	
		+	-
PREDICTED	+	**TP**	**FP**
	-	**FN**	**TN**

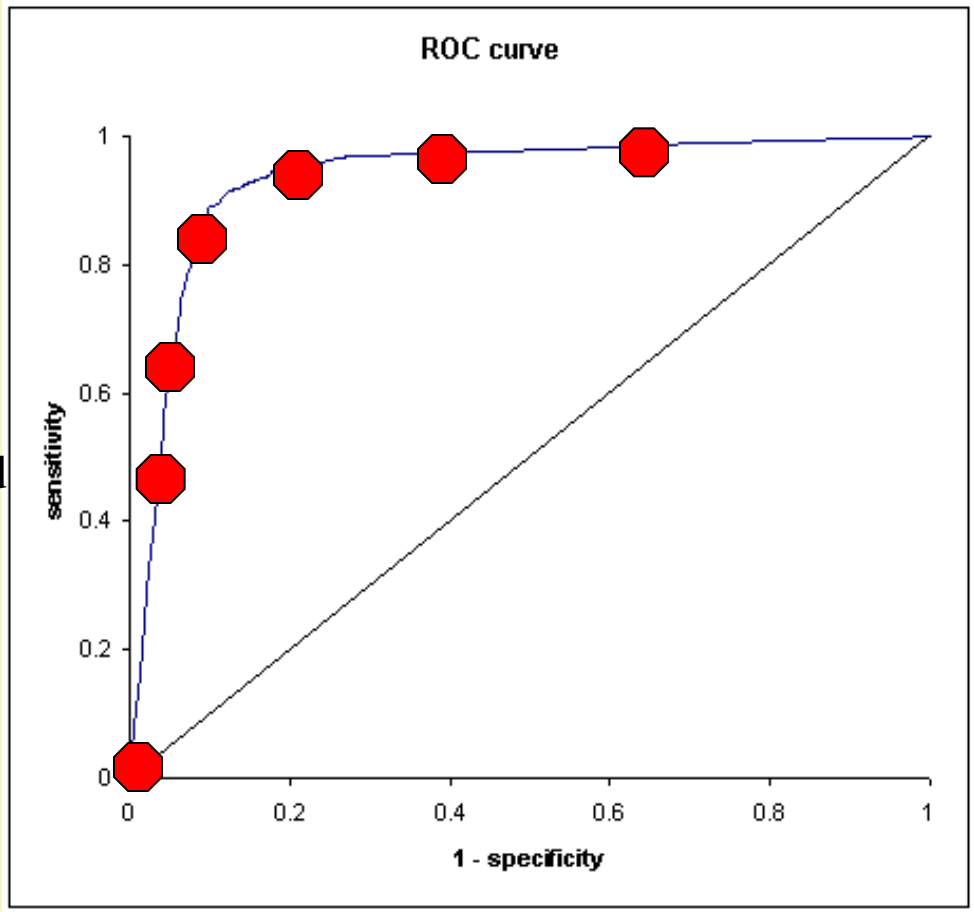
Sensitivity = TP / All Actual Positives

Specificity = TN / All Actual Negatives

A Reciever Operator Characteristic (ROC) plot

sensitivity = a/(a+c)

Fraction of presences predicted present.

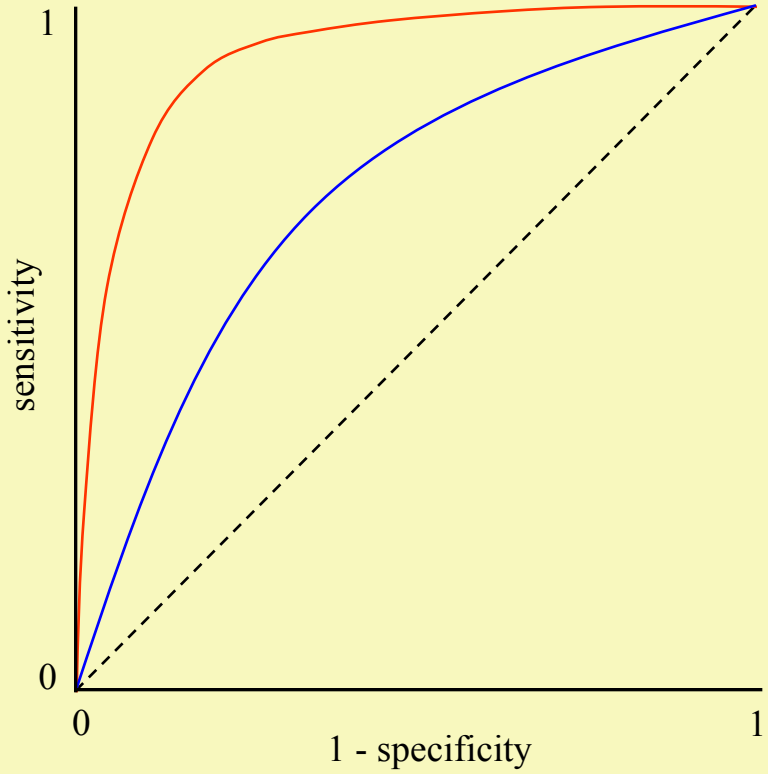


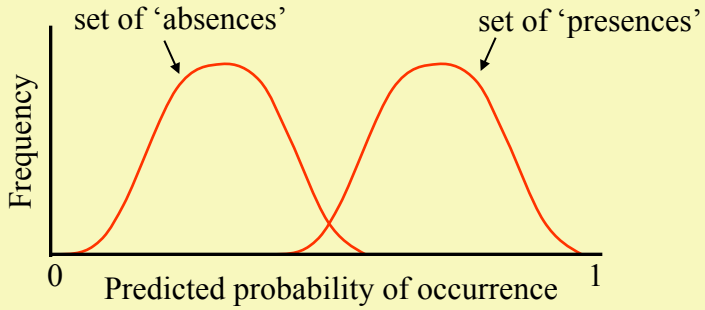
1- specificity = 1 – [d/(b+d)]

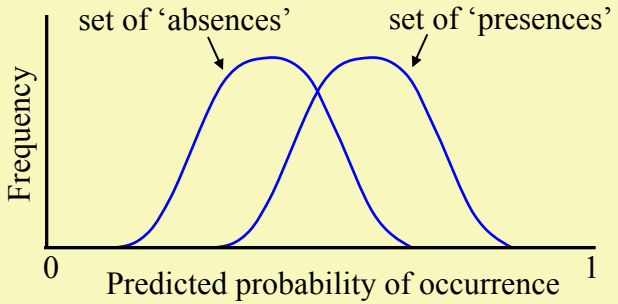
Fraction of absences predicted present

	Recorded present	*Recorded (or assumed) absent*
Predicted present	a (true positive)	b (false positive)
Predicted absent	c (false negative)	d (true negative)

Threshold-independent assessment:
The Receiver Operating Characteristic (ROC) Curve







(check out: http://www.anaesthetist.com/mnm/stats/roc/Findex.htm)

What is a ‘good’ result?

Some *subjective* guidelines:

Kappa (after Landis & Koch 1977 *Biometrics*):

- 0 – 0.4: poor
- 0.4 – 0.75: good
- 0.75 – 1.0: excellent

AUC (after Swets 1988 *Science*):

- 0.5 – 0.7: poor discrimination
- 0.7 – 0.9: reasonable discrimination
- 0.9 – 1.0: very good discrimination