

# Algunas reflexiones sobre tipología de modelos con motivo de nichos ecológicos

Enero 2007

Dr. Miguel Nakamura Savoy  
Centro de Investigación en Matemáticas, A.C.

Guanajuato, Gto.

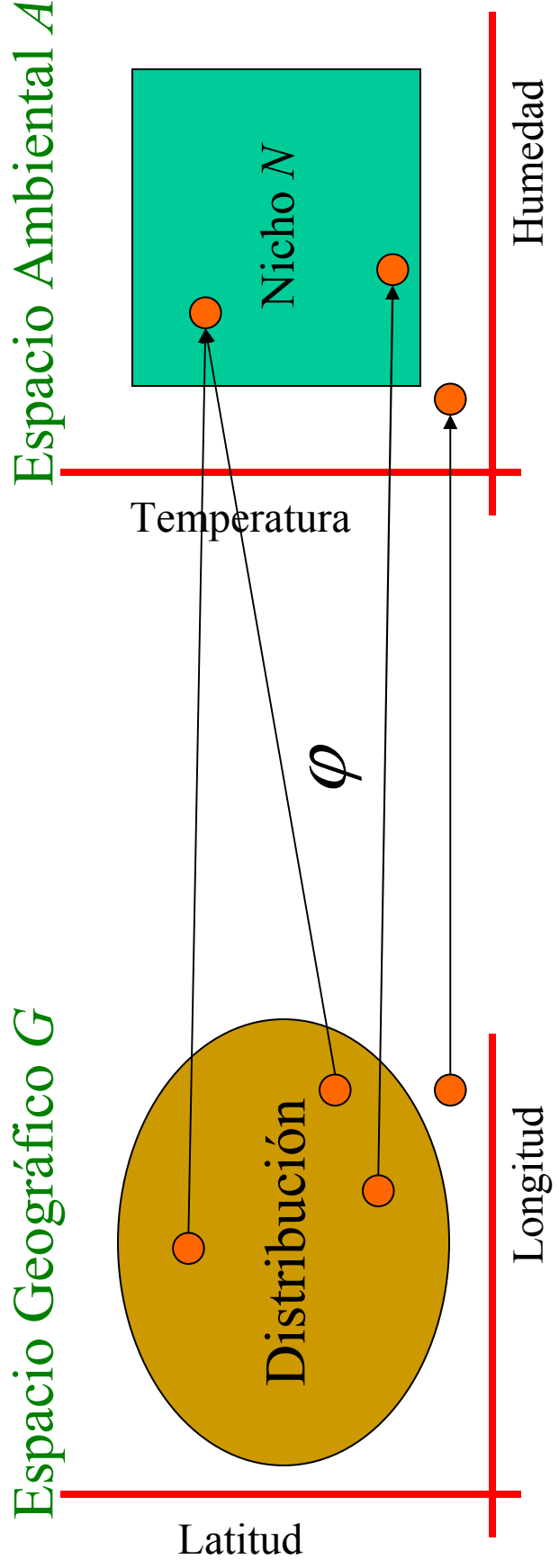


# Nichos Ecológicos



# Biología: Nicho Ecológico

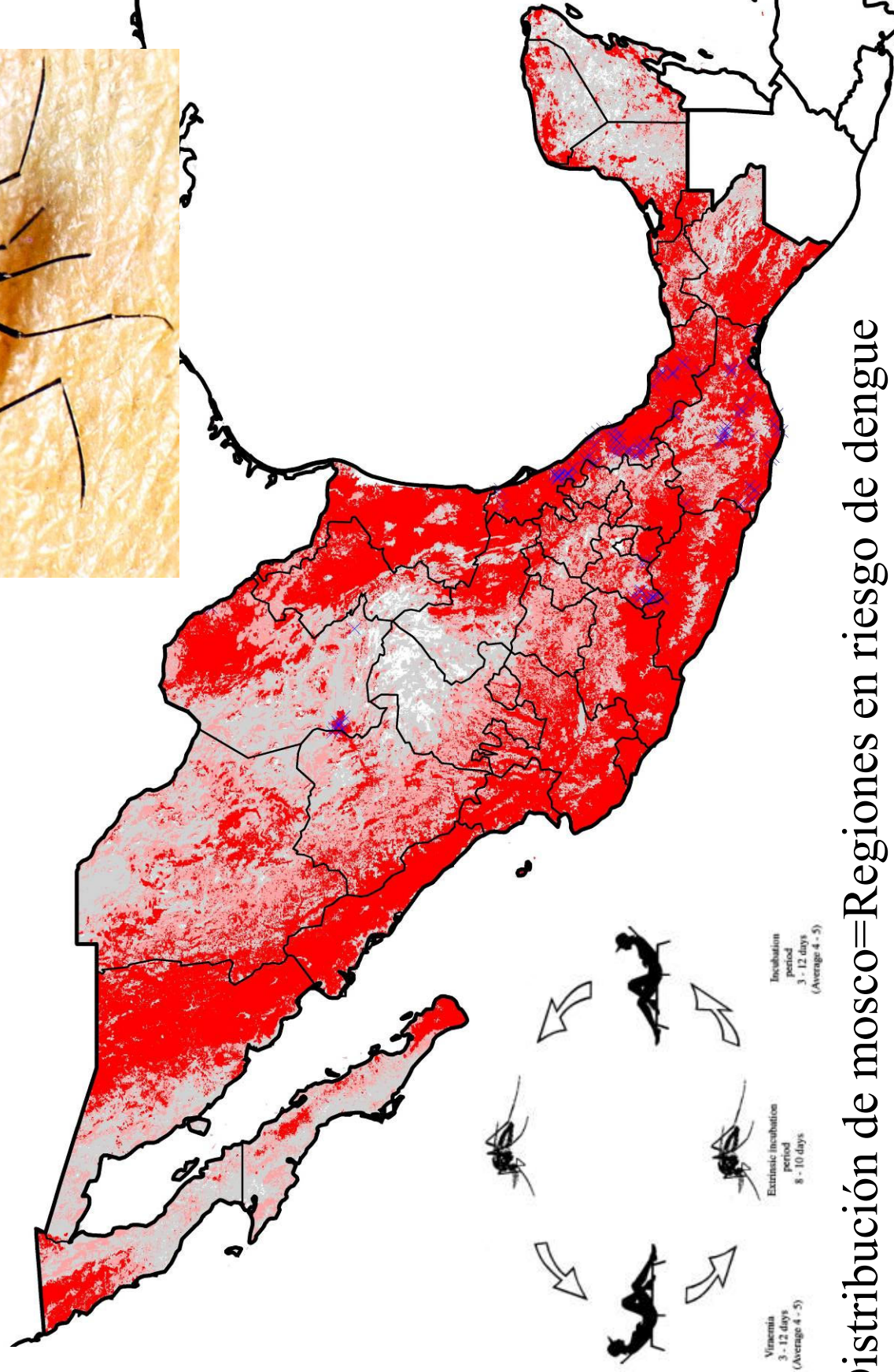
- The particular area within a habitat occupied by an organism.
- “...the ultimate distributional unit, within which each species is held by its structural and instinctive limitations” (Grinnell, 1924).
- “The term niche...is here defined as the sum of all the environmental factors acting on the organism; the niche thus defined is a region of an n-dimensional hyper-space...” (Hutchinson, 1944).
- “The niche of a species is the joint description of the environmental conditions that allow a species to satisfy its minimum requirements so that the birth rate of a local population is equal or greater than its death rate along with the set of per capita impacts of that species on these environmental conditions” (Chase and Leibold, 2003).



Ambiente  $\varphi : G \rightarrow A$

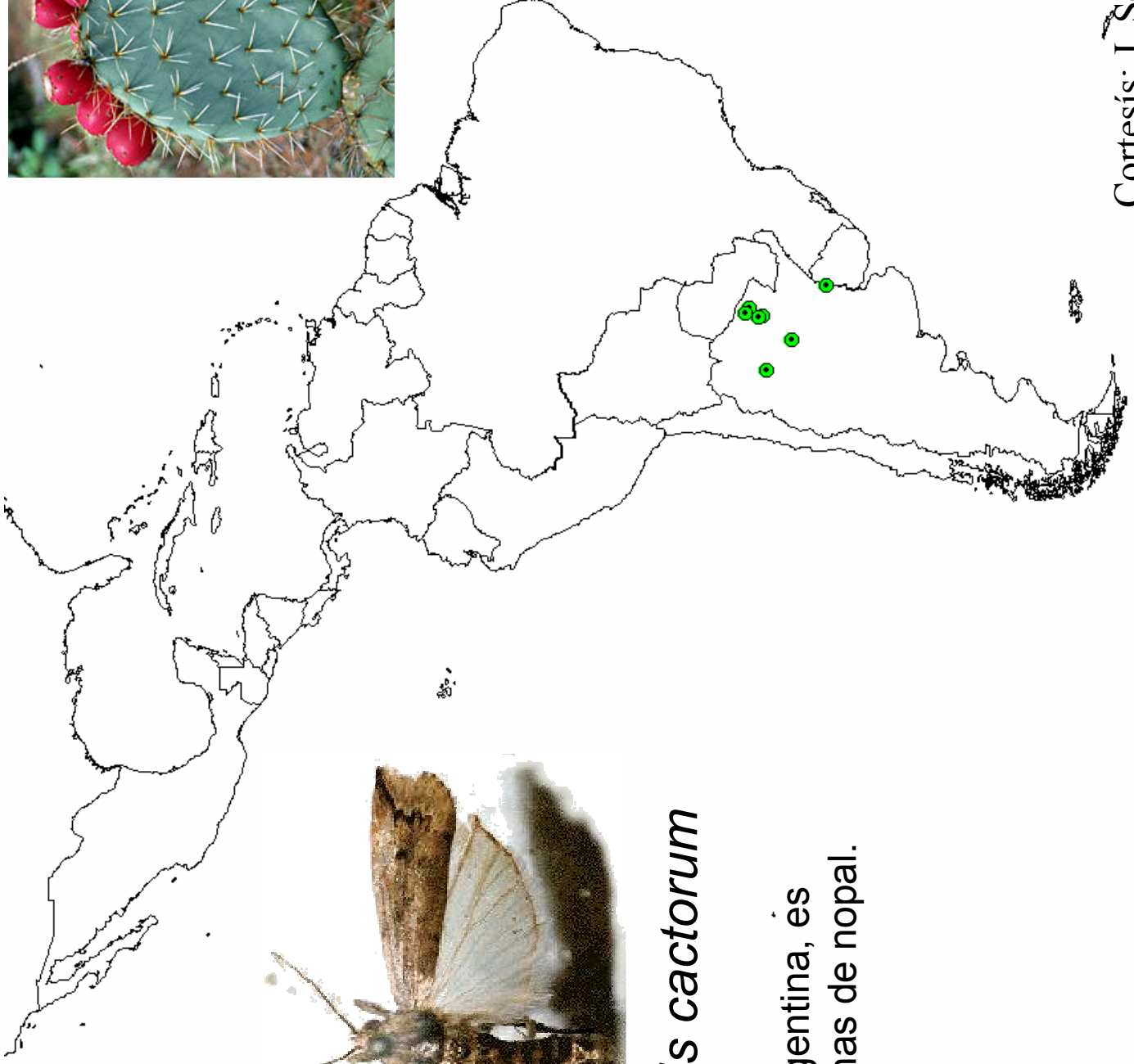
"Distribución":  $\varphi^{-1}(N) \subset G$

**Mosco del dengue: *Aedes aegypti***



Distribución de mosco=Regiones en riesgo de dengue

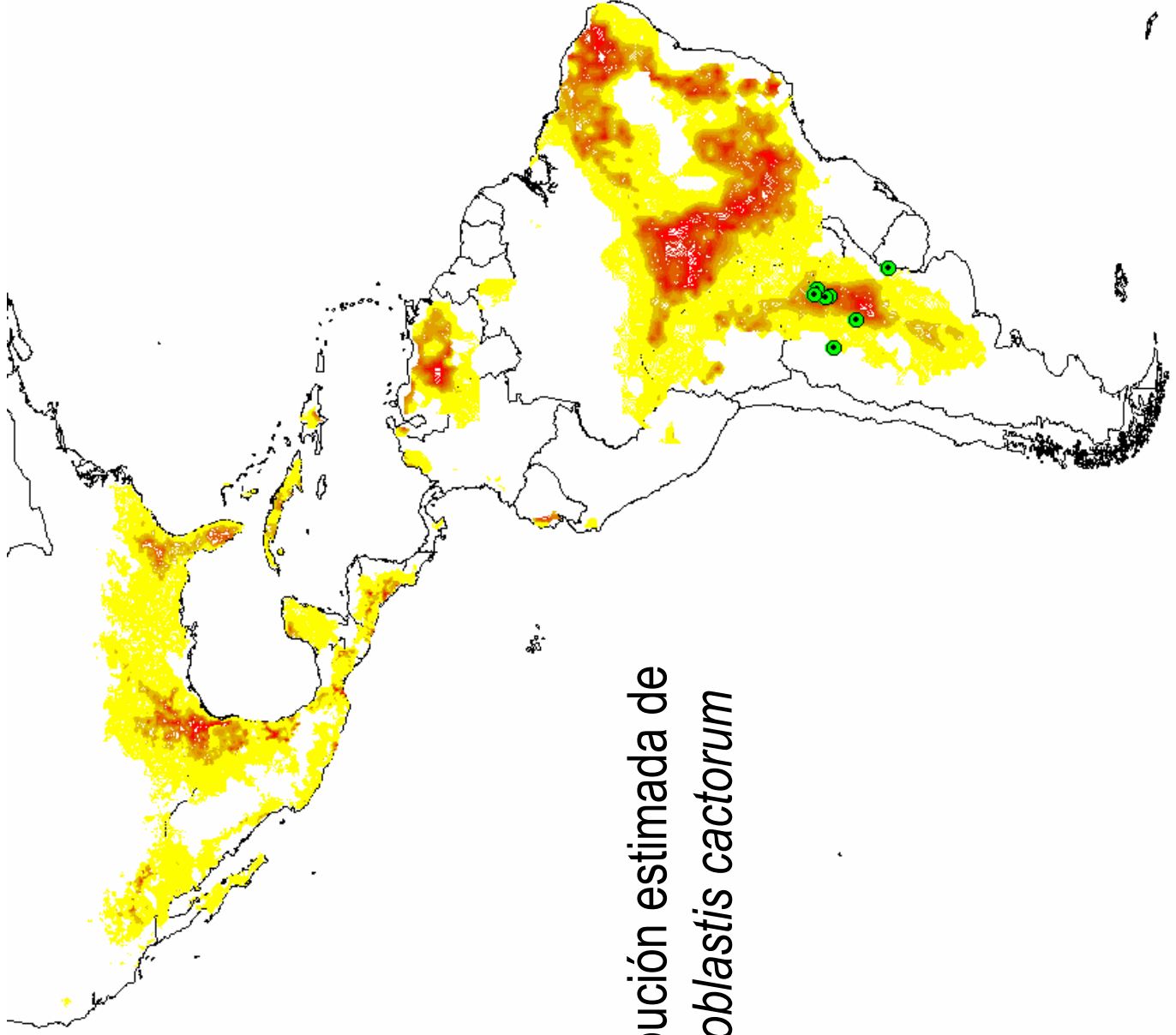
Cortesía: Y. Nakazawa



## *Cactoblastis cactorum*

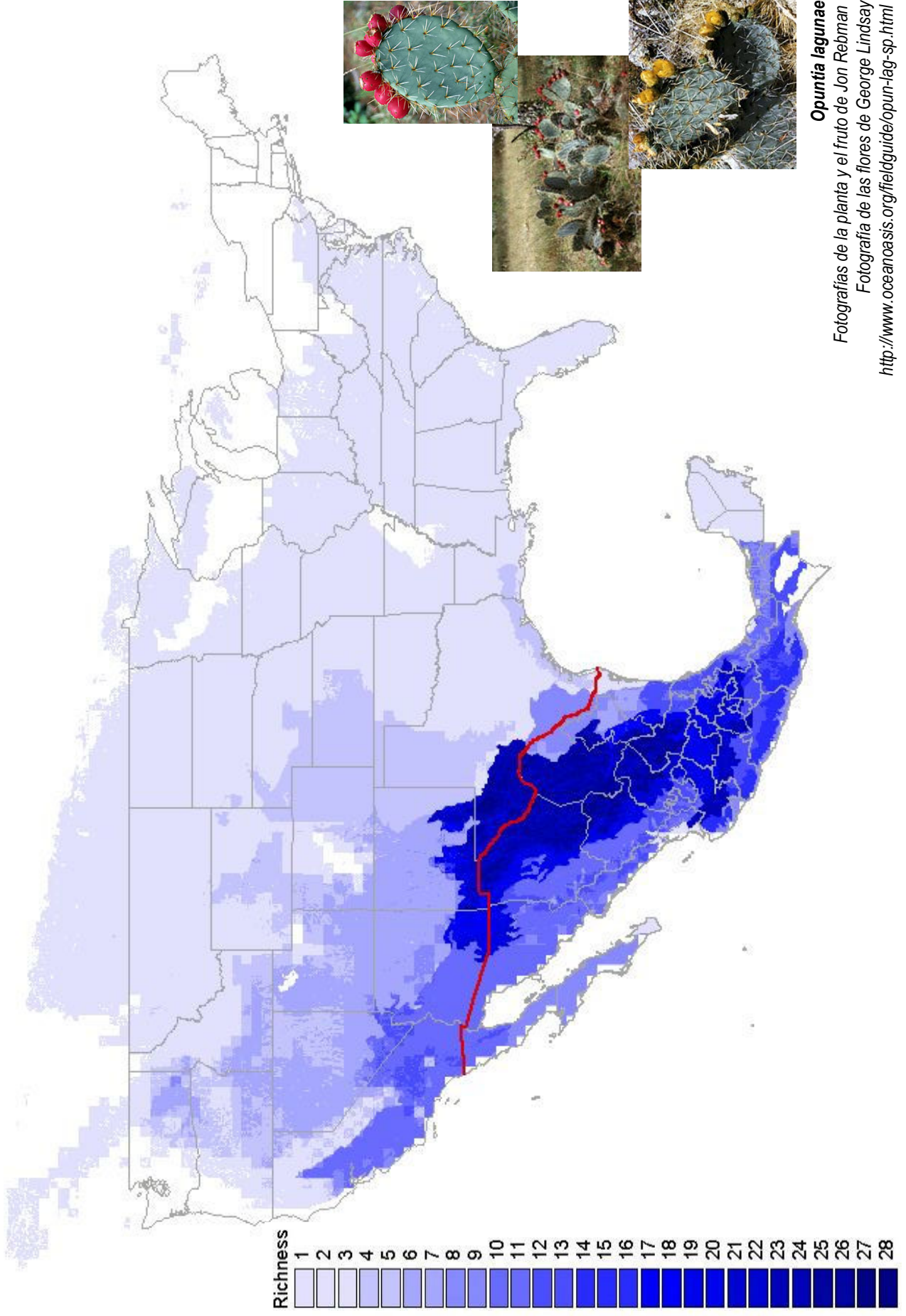
Originario de Argentina, es devorador de tunas de nopal.

Cortés: J. Soberón



Distribución estimada de  
*Cactoblastis cactorum*

# Riqueza de especies de *Platyopuntia*



***Opuntia lagunae***  
Fotografías de la planta y el fruto de Jon Rebman  
Fotografía de las flores de George Lindsay  
<http://www.oceanoasis.org/fieldguide/opun-lag-sp.html>

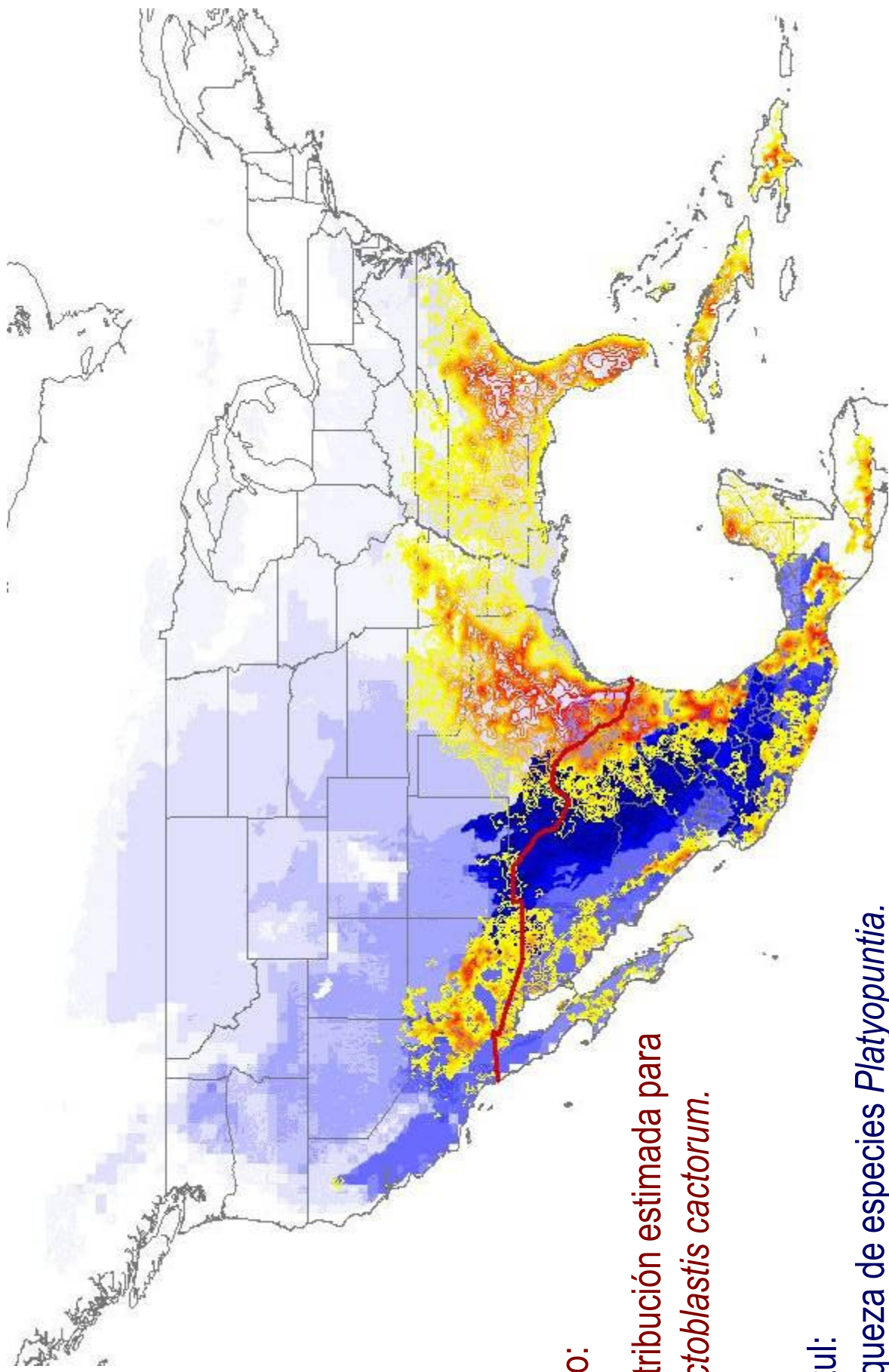


Tunas



Ensalada de nopalitos

Areas vulnerables a *Cactoblastis* (clima/alimento apropiado)



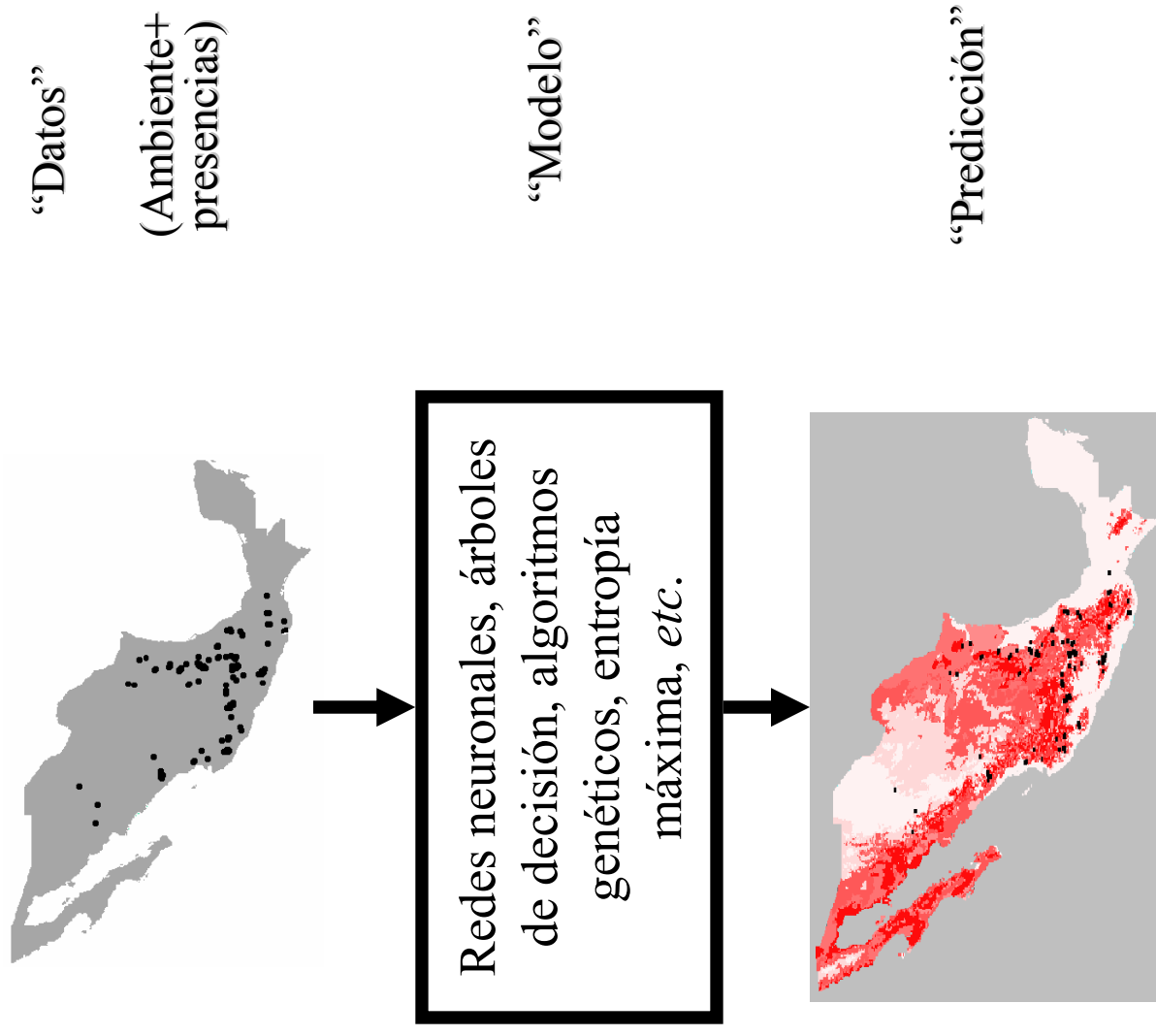
Rojo:  
Distribución estimada para  
*Cactoblastis cactorum*.

Azul:  
Riqueza de especies *Platyopuntia*.

## Por qué los nichos y distribuciones son científicamente relevantes:

- Salud pública.
- Predicción de especies invasoras.
- Planeación de conservación.
- Estudios de evolución.
- Predicción de migración por cambio climático.
- Acelerar descubrimiento de nuevas especies.
- ... *etc.*

# ¿Cómo se obtienen distribuciones estimadas?



# ¿Qué hace que un método sea estadístico?

- Para lego en la materia, el carácter estadístico viene por el hecho de que se usan datos observados empíricamente como entrada.
- La profesión estadística tiende a definirlo en términos de las herramientas que se usan (e.g. modelos de probabilidad, Cadenas de Markov, ajuste por mínimos cuadrados, teoría de verosimilitud, etc.)
  - Ejemplo: “This chapter is divided into two parts. The first part deals with methods for finding estimators, and the second part deals with evaluating these (and other) estimators.” (Casella & Berger, *Statistical Inference*, 1990)
- Algunos autores sugieren un concepto mucho más amplio de estadística, con base en el tipo de problemas que la estadística pretende resolver.
  - Ejemplo: “Perhaps the subject of statistics ought to be defined in terms of problems, problems that pertain to analysis of data, instead of methods”. (Friedman, 1977)

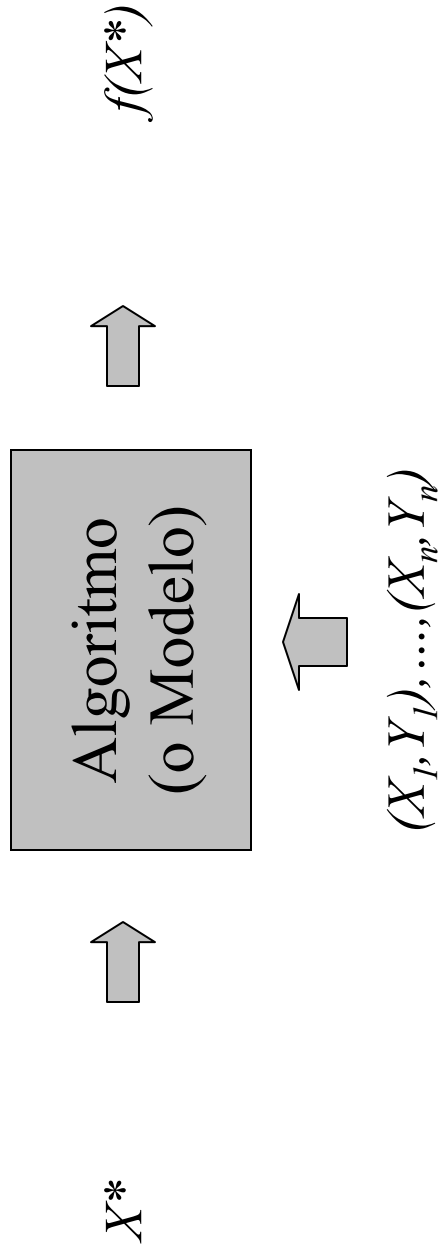
# Entonces, es el problema el que es estadístico, no el modelo.

- Un problema estadístico se caracteriza por: datos sujetos a variación, una pregunta de interés, incertidumbre en la respuesta que puede obtenerse con los datos, algún grado de razonamiento inferencial.
- ¿Por qué razón la profesión estadística enfatiza ciertos tipos de métodos? Porque la variación en los problemas estadísticos se reconoce de entrada, y la cuantificación de incertidumbre se toma por hecho y es rutinariamente abordada por tales métodos. Así, uno podría entender como “método estadístico” en el sentido de cuantificación de incertidumbre. En este sentido algunos estadísticos no considerarían algunos análisis de datos como análisis “estadísticos”.

## ¿En principio, problema binario de clasificación?

- $X$  es vector de variables ambientales.  $Y=1, 0$  es presencia ausencia.
- Hay una función  $f(X)$  (desconocida) que describe probabilidad condicional de presencia vs. ausencia.
- Se cuenta con una muestra empírica de ambientes ( $X$ 's) asociados con presencias y con ausencias ( $Y$ 's) (?).
- El objetivo es predecir  $Y^*$ , con base en su  $X^*$ , si un lugar arbitrario pertenece a la distribución o si no pertenece, es decir,  $f(X^*)$ . (Quizás si  $f(X^*) > 1/2$ , se declara una presencia).

# Algoritmos vistos como “cajas negras” (en nichos ecológicos)



## Ejemplos de algoritmos

- Basados en envolventes (BIOCLIM)
- Basados en distancias (LIVES, DOMAIN)
- Basados en regresión (Modelos aditivos generalizados, lineales generalizados, regresión logística).
- Redes neuronales.
- Boosted regression trees.
- Algoritmos genéticos (GARP, Stockwell & Peters, 1999).
- Entropía máxima (Maxent, Phillips *et al.*, 2006).
- Otros (ver Elith, *et al.*, 2005)

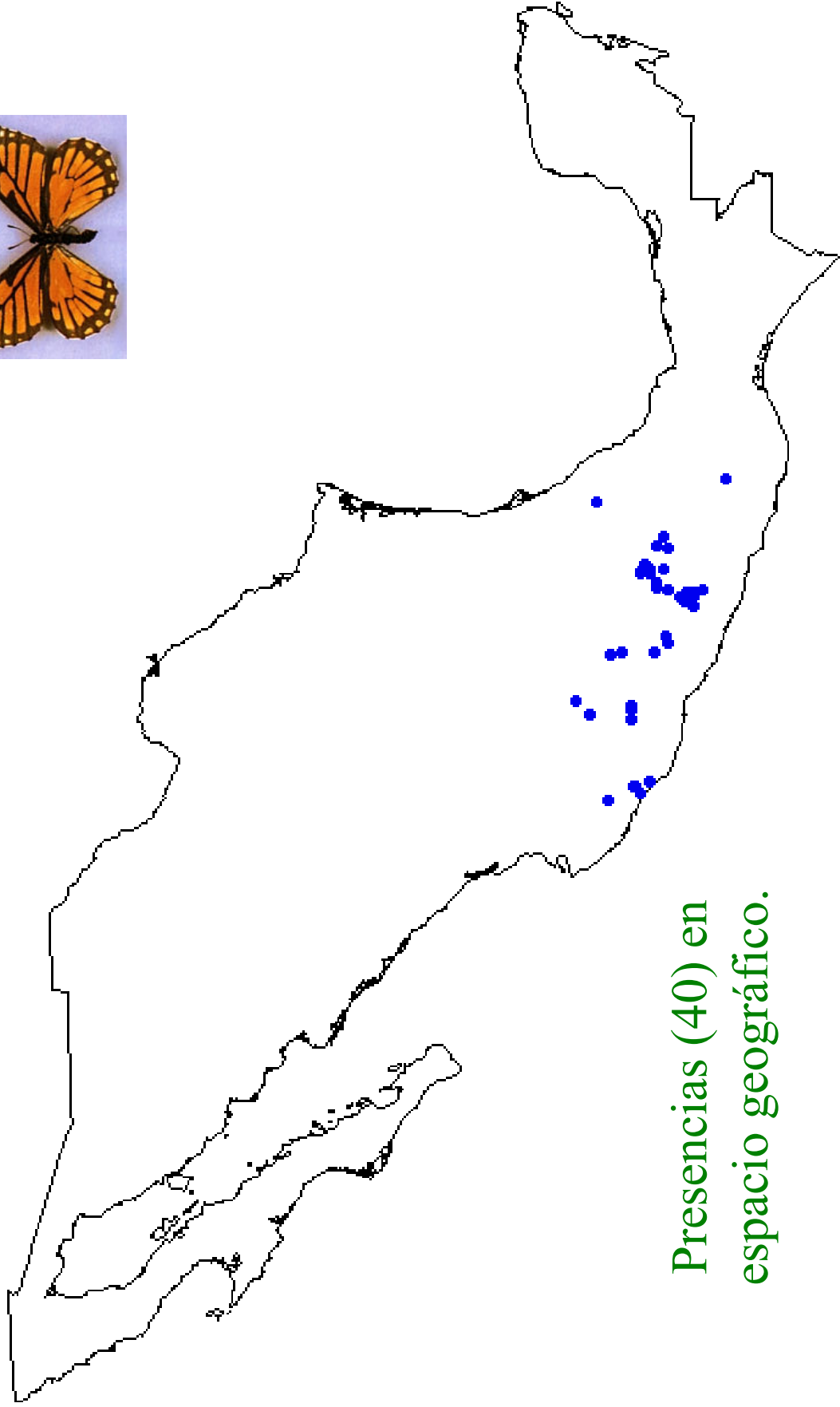
## Un ejemplo

### *Baronia brevicornis*

Variables ambientales: clima, humedad, tipo de suelo, precipitación, temperatura media, temperatura máxima, temperatura mínima, elevación.



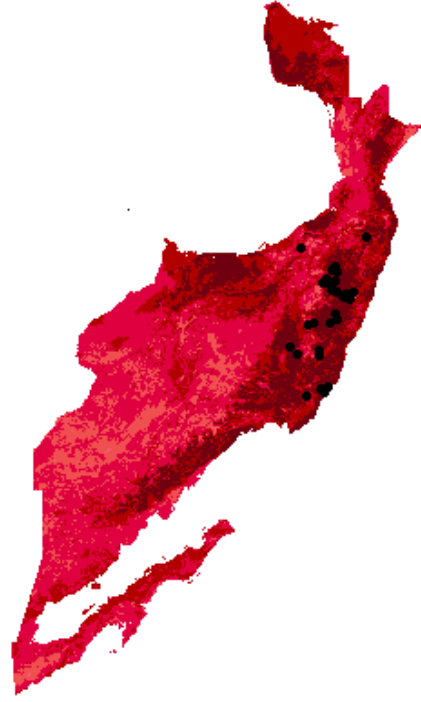
*Baronia brevicornis*



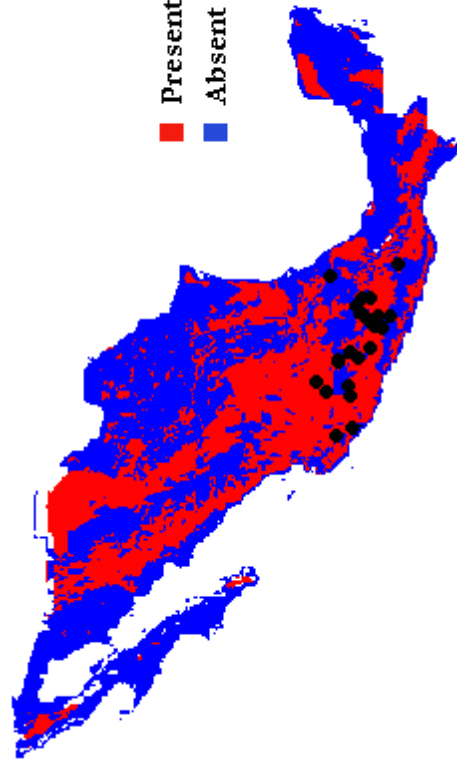
Presencias (40) en  
espacio geográfico.

# Baronia brevicornis: Distribuciones predichas

Domain

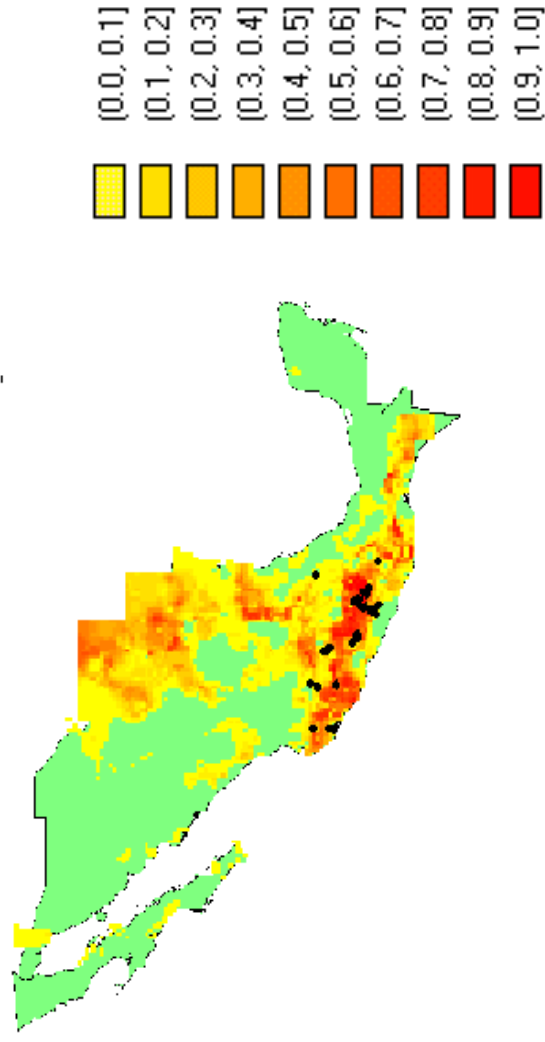


GARP



Present  
Absent

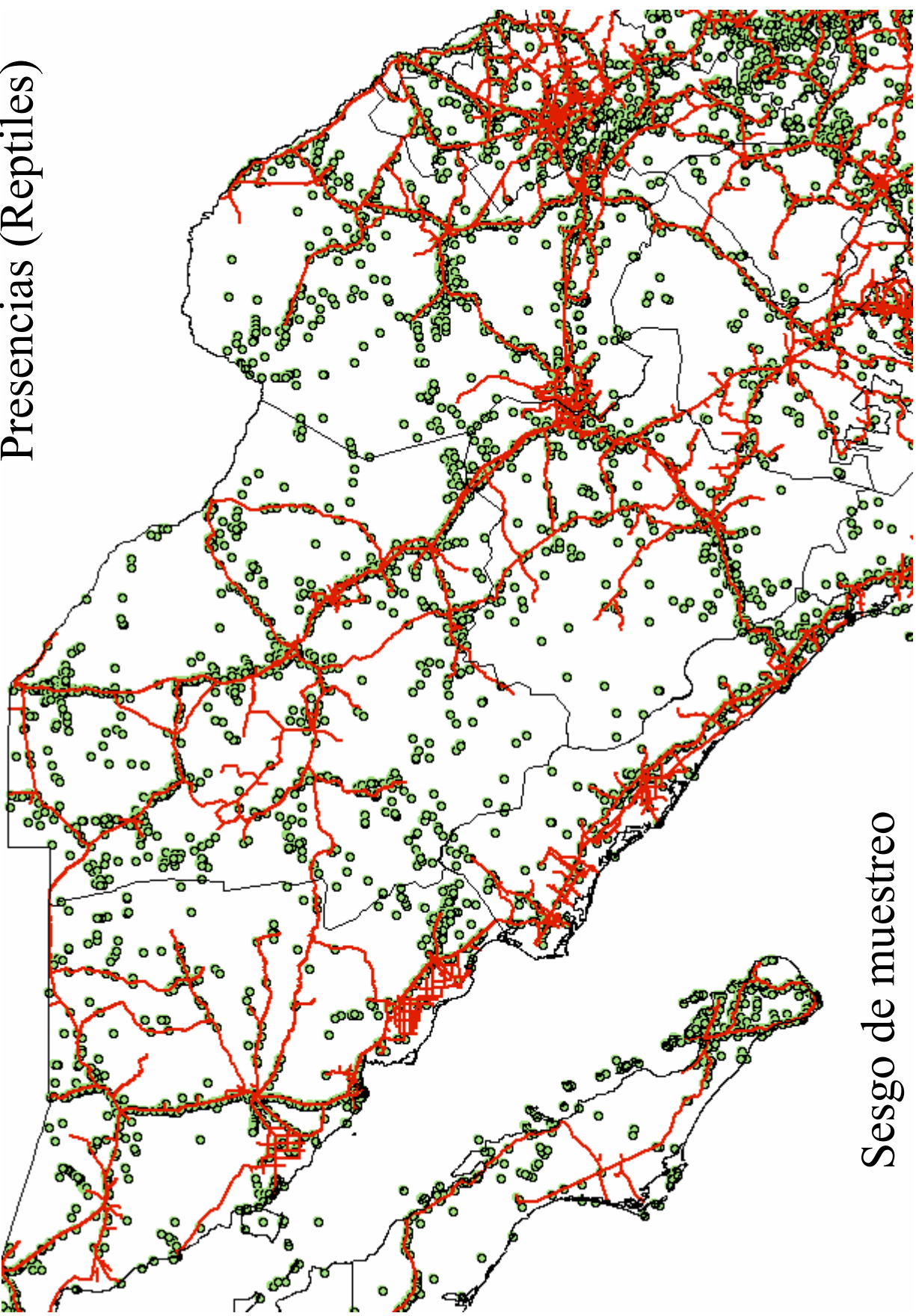
FloraMap



## Algunos detalles propios al contexto de nichos ecológicos

- ¡Ausencias no son ausencias legítimas!
- La mayor parte de las veces sólo hay presencias, por lo que se generan “pseudoausencias”.
- No son datos obtenidos bajo diseño, sino incidentales. El muestreo no es homogéneo, ni probabilístico.
- Existe información *a priori*.
- Hay problemas de escala.
- Hay correlación espacial.
- Deficiencia en cantidad y calidad de datos.
- Alta dimensionalidad (en espacio ambiental).
- Objetivos múltiples.
- Consideraciones biológicas son extremadamente complejas.
- Hay escenarios muy numerosos (número de especies, número de regiones).

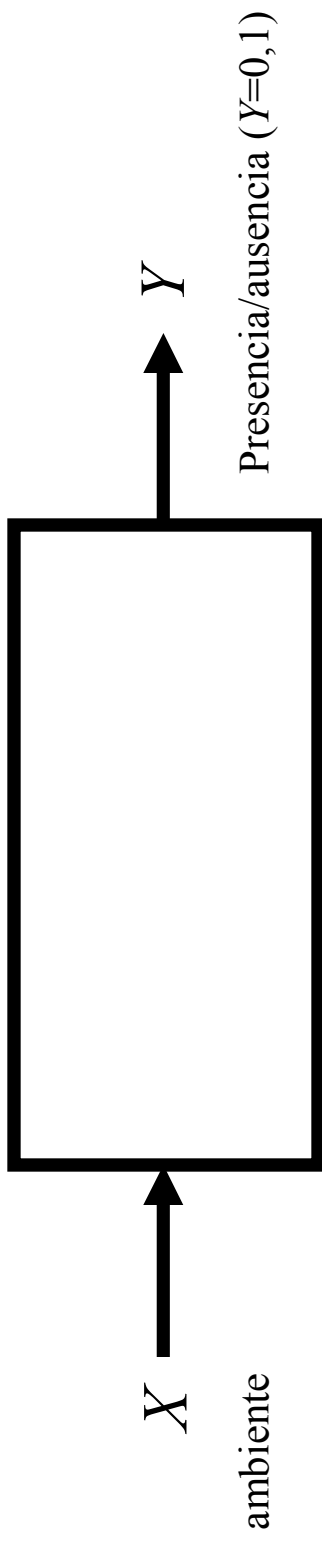
Presencias (Reptiles)



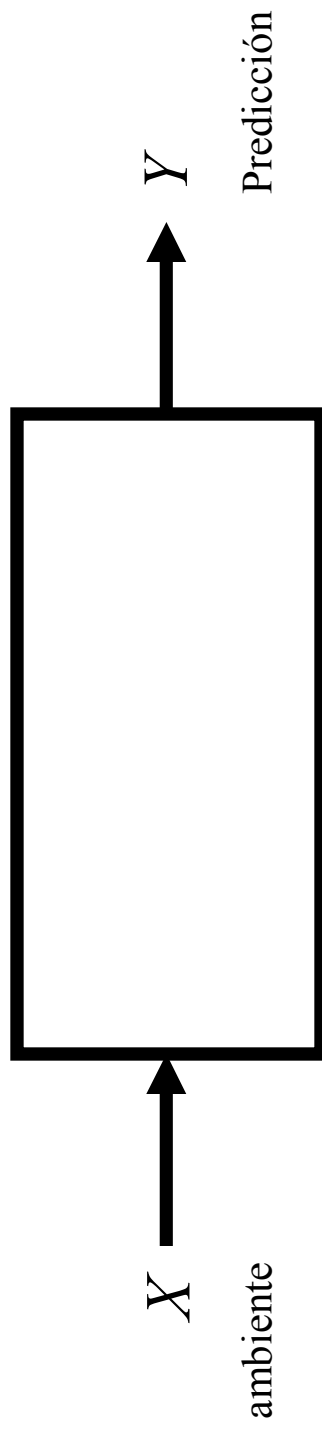
Sesgo de muestreo

# Hacia una taxonomía de modelos

Caja negra de la naturaleza



Caja negra del modelador

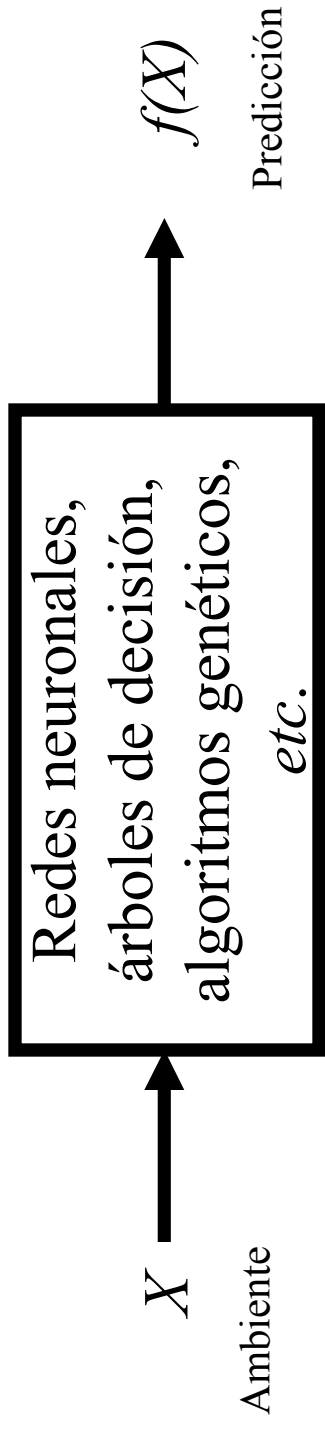


# La caja negra del modelador: Las dos culturas\*

1. Algorithmic Modeling (AM) culture
2. Data Modeling (DM) culture

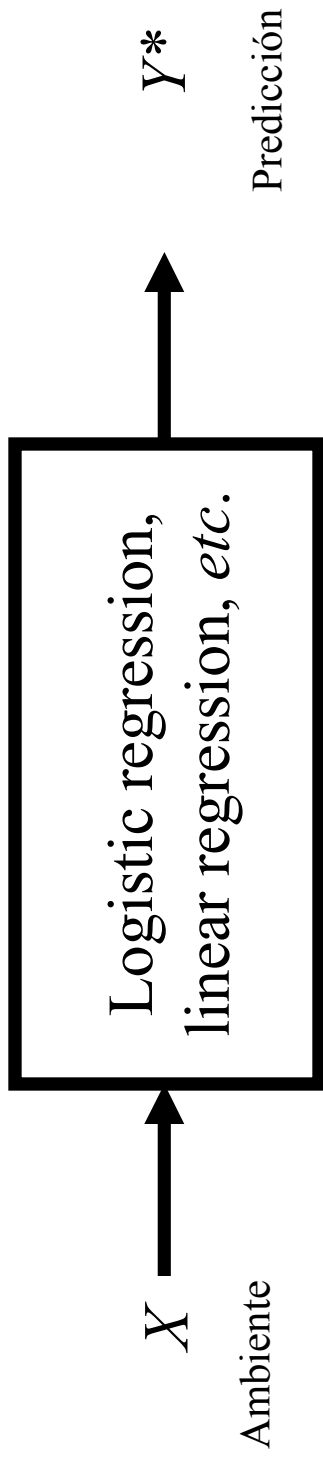
\* Breiman (2001), *Statistical Science*, with discussion.

# Cultura de modelación algorítmica



- El interior de la caja negra es complejo y desconocido. La interpretación es usualmente difícil.
- El enfoque es encontrar una función  $f(x)$  (un algoritmo).
- Validación: examinar precisión predictiva.
- La noción de incertidumbre no necesariamente es considerada.

# Cultura de modelación de datos



- Hay un modelo probabilístico dentro de la caja negra. Esto implica suposiciones acerca de la aleatoriedad.
- $P(Y)$  (modelo de probabilidad para  $Y$ ) es producto secundario (el cual a la vez permite hacer aseveraciones para cuantificar la precisión de la predicción).
- Hay parámetros que se estiman vía datos observados.
- El método de predicción es prescrito por el modelo y/o por objetivos y suposiciones (también está dentro de la caja negra).
- Validación: análisis para determinar si las suposiciones sobre aleatoriedad son válidas. Término “bondad de ajuste”.

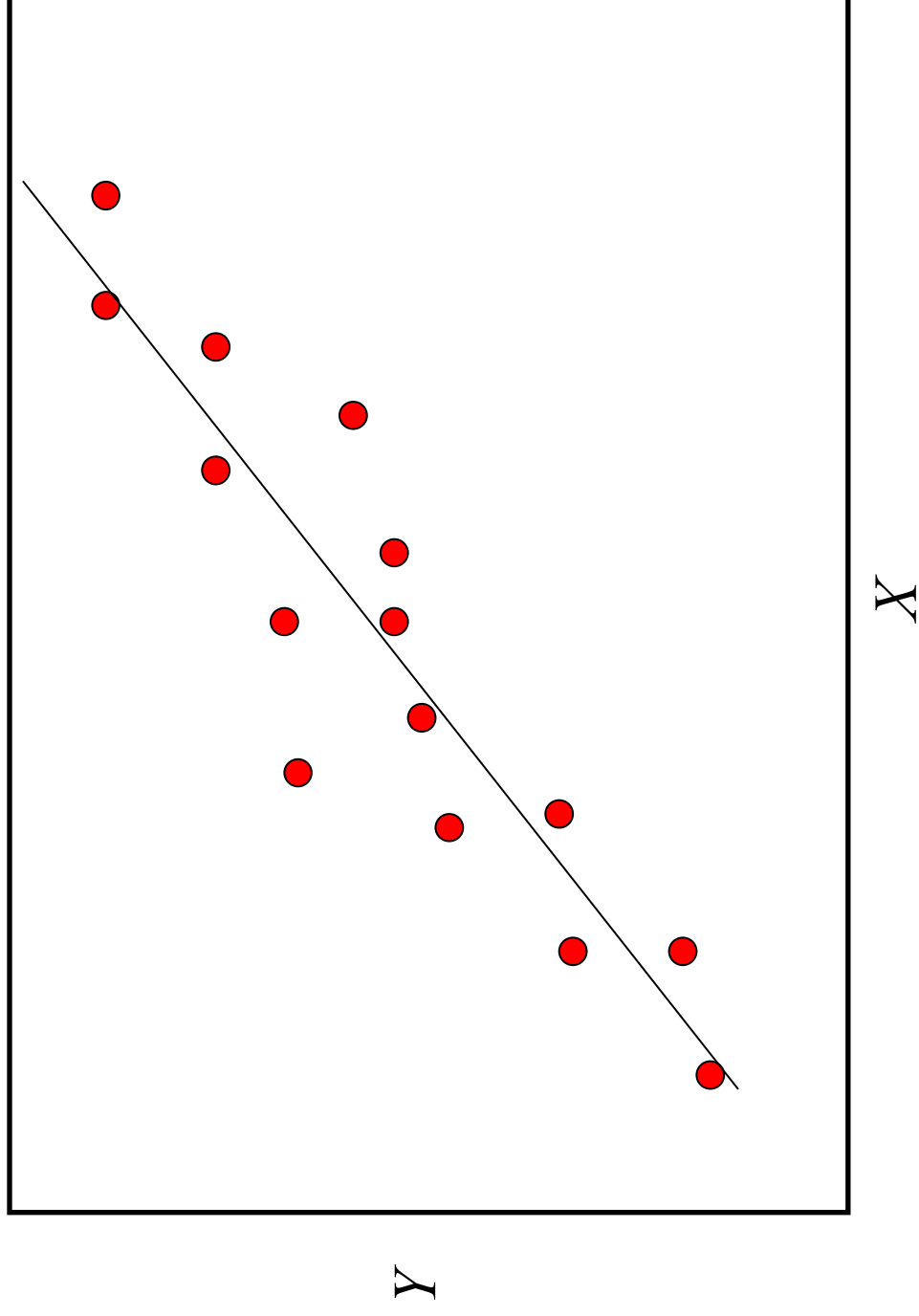
## Taxonomía de modelos de probabilidad en la cultura de modelación de datos: subespecies\*

- *Modelos sustantivos*, que se basan en conocimiento de la materia y contexto, un mecanismo que explica lo que es observado.
- *Modelos empíricos*, que quieren representar de una manera “suave” el comportamiento a largo plazo de repeticiones, sin basarse en un contexto específico.

\* Cox (1990), *Statistical Science*

# Diferencia entre algoritmo y modelación de datos.

Ejemplo ilustrativo: Regresión lineal simple



## Punto de vista modelación de datos

- Se asume que cada observación de  $Y$  tiene una distribución de probabilidad (e.g. normal) para cada  $X$ . La estructura lineal es una suposición que puede provenir de conocimiento en la materia (e.g. teoría química). Una consideración de índole probabilístico (máxima verosimilitud) produce el ajuste por mínimo cuadrados.
- Como parte del proceso de ajuste, se obtiene la cuantificación de incertidumbre en los parámetros estimados (intercepto y pendiente).

# Punto de vista modelación algorítmica

- No hay un papel explícito para probabilidad. Los puntos  $(X, Y)$  se aproximan por una línea recta. Un argumento geométrico (no-probabilístico) de minimización de una distancia también da lugar a un ajuste por mínimos cuadrados. El intercepto y pendiente no necesariamente interpretables, ni de interés especial.
- El modelo de datos y el enfoque algorítmico ambos dan lugar a un ajuste por mínimos cuadrados. ¿Esto significa que ambos están haciendo lo mismo y persiguiendo los mismos objetivos? ¡No! Para modelo de datos, la recta es un estimador de un concepto probabilístico. Para modelo algorítmico, la recta es un dispositivo de aproximación.
- Si del análisis estadístico sólo se extrae la recta ajustada, se ignora por completo la descripción de variabilidad de  $Y$  (por vía del modelo de probabilidad) que está presente en el modelo.

# Explanation or prediction? Inference or decision?

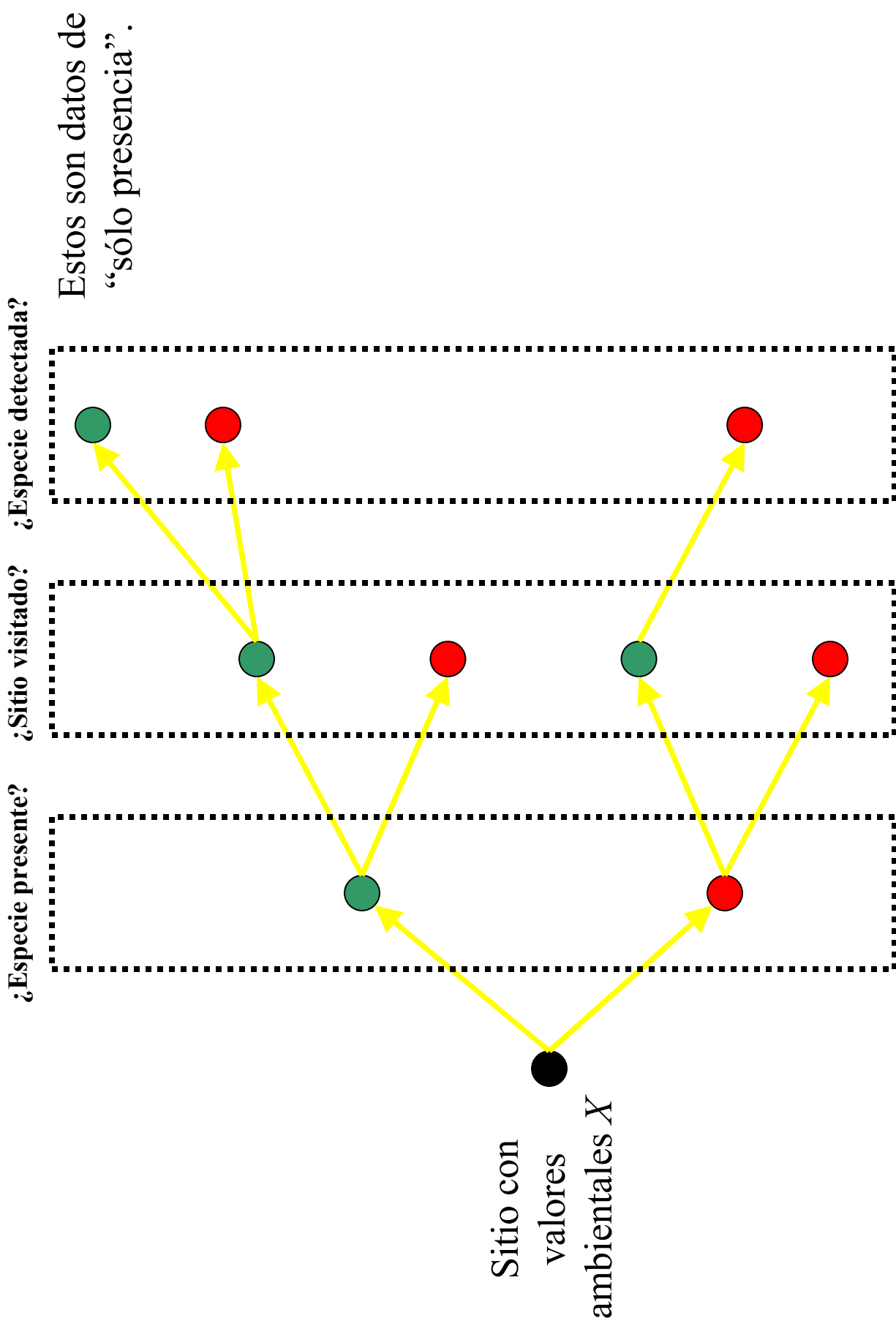
- For explanation, Occam's razor applies. A working model, a sufficiently good approximation that is simple is preferred.
- For prediction, all that is important is that it works. Modern hardware+software+data base management have spawned methods from fields of artificial intelligence, machine learning, pattern recognition, and data visualization.
- Depending on particular need, some models may not provide the required answers.

# Ejemplo en cultura de modelación de datos\*:

## La relación entre datos y el nicho

- Al menos los siguientes conceptos parecerían ser relevantes dado el contexto del problema:
  - Sesgo espacial de muestreo.
  - Detectabilidad.
- Existe aleatoriedad en el proceso de observación.
- Existe información (biológica) previa legítima.
- El objeto de interés es probabilidad de presencia en un sitio. Se requiere inferencia estadística para probabilidad de presencia:
  - Estimaciones.
  - Precisión de estimaciones.

\* [Argáez, Christen, Nakamura, and Soberón \(2005\)](#)



Una presencia es el resultado de este mecanismo aleatorio.

# Ejemplo ilustrativo:

100 ensayos de muestreo

Especie #1

Clima A      Clima B      Clima C

Presencia                 

×      ×      ×

Sesgo                 

=      =      =

Tasa de obs.                 

Observados                 

Especie #2

Clima A      Clima B      Clima C

×      ×      ×

=      =      =

100–16 – 8=76 ensayos no observados; 24 presencias observadas en ambos casos.

## Aspecto ilustrado por el ejemplo

- Como los datos observados son idénticos, cualquier método que sólo utilice esos datos observados (presencias en climas A,B,C), es incapaz de discernir entre Especie #1 y Especie #2.
- Sesgo de muestreo y otras condiciones se vuelven cruciales. Sólo se identifican cuando se examina la génesis de los datos. Esta es la cultura de modelación de datos.

# Modelación Probabilística de Datos

Sitio  $s$  tiene variables ambientales  $X$ .

$$P(\text{observar en } X) = P(\text{presencia en } X) \times P(X \text{ is visitado}) \times P(\text{detectado})$$



**Datos**



**Probabilidad de presencia.**

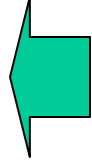
Es función de  $X$ . Este es parámetro de interés.

Modelación de probabilidad de presencia.



**Sesgo en muestreo.**

Es función conocida de  $X$  y sesgo espacial, ésta supuesta dada.



**Probabilidad de detección.**

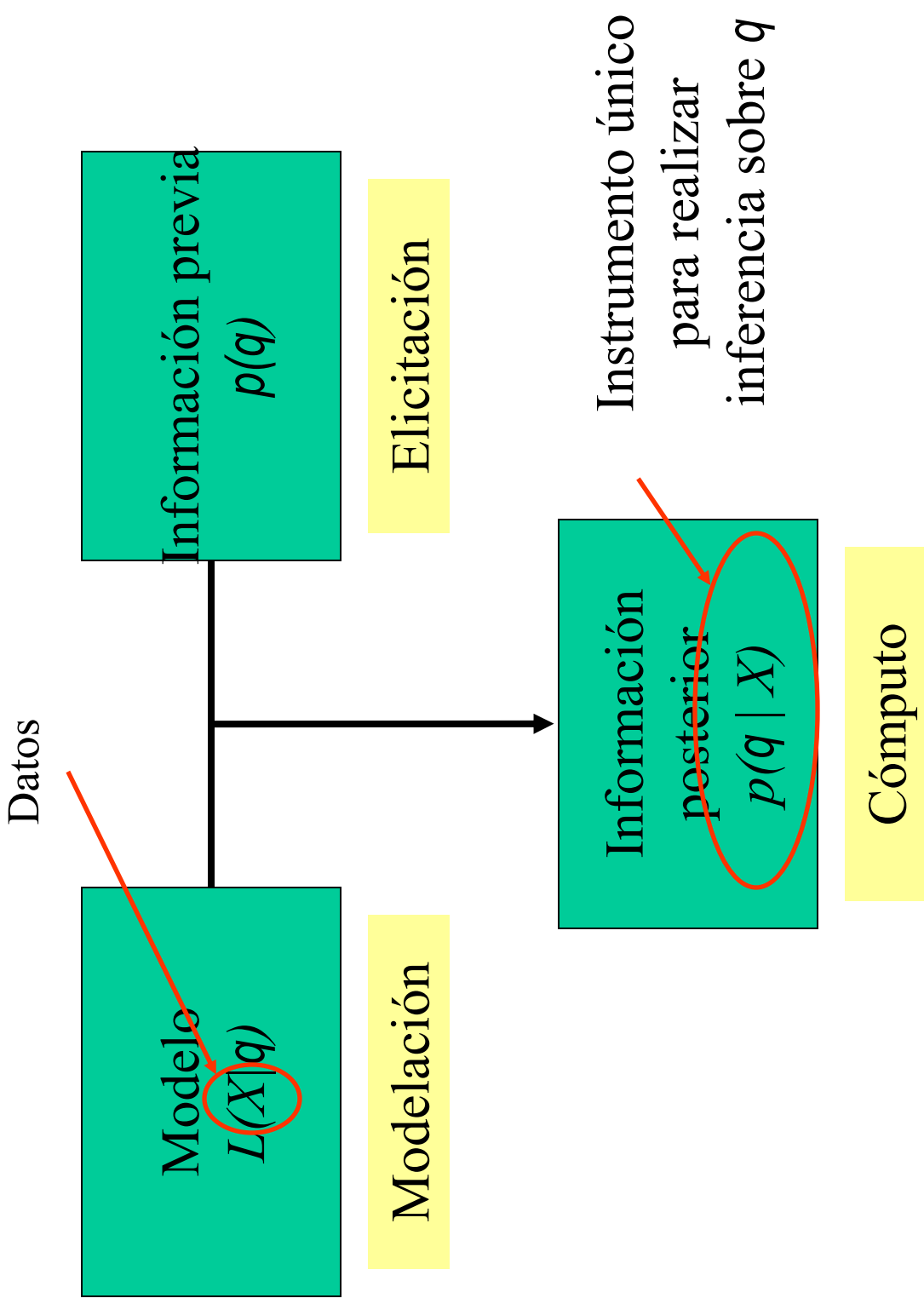
Se supone dado.

# ¿Qué otras cosas afectan la distribución de los datos\*?

- Una cosa es distribución de la especie, y otra es distribución de los datos. Debe entenderse la relación entre datos y nicho.
- El diagrama de árbol previo es más complejo:
  - Accesibilidad.
  - Interacciones.
  - Poblaciones sumidero.
  - Resolución de la rejilla.
  - Nicho realizado vs. potencial.
- Algunos casos especiales permiten simplificaciones:
  - Muestreo uniforme.
  - Detección constante.
  - Accesibilidad irrestricta.

\* Nakamura, and Soberón (2007?)

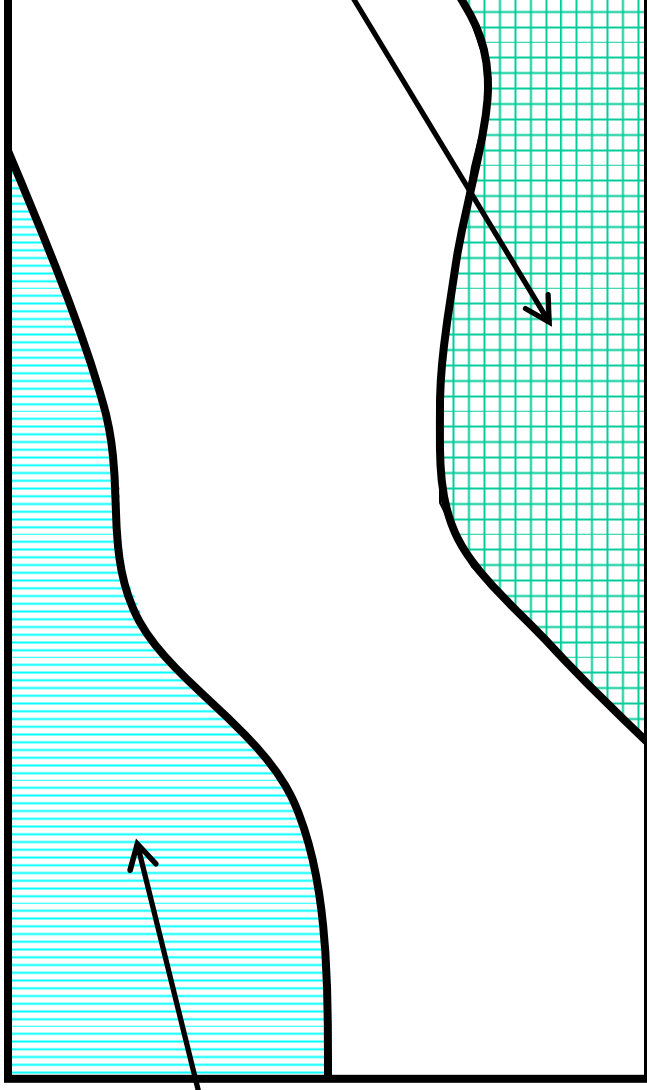
# Inferencia Estadística: Paradigma bayesiano



# Elicitación

**Convertir esto:**

Usuario  
especifica  
especie es  
muy  
probable sí  
se encuentre  
aquí.



Usuario  
especifica  
especie es  
muy  
improbable  
se  
encuentre  
aquí.

**a esto:**  $p(q)$ , densidad de probabilidad que cuantifica conocimiento sobre  $q$

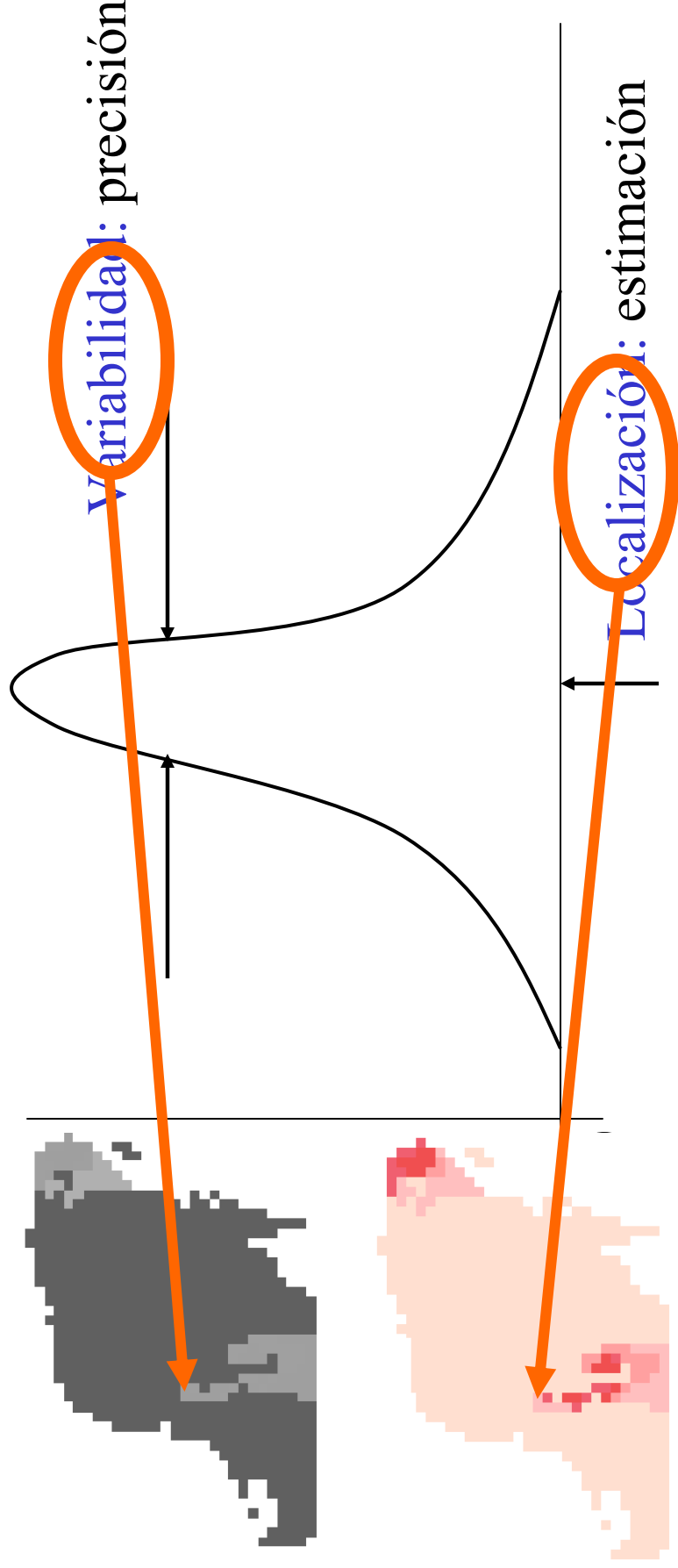
**Herramientas:** Familia Dirichlet, noción de información contradictoria.

# Cómputo

Calcular  $p(q | X)$ , densidad posterior

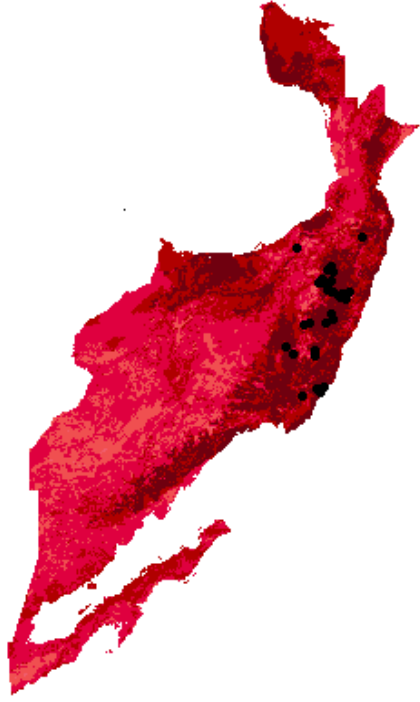
**Herramientas:** MCMC (Markov Chain Monte Carlo), aproximaciones numéricas.

**Output:** una densidad de probabilidad para cada  $P$  (presencia en  $X$ )

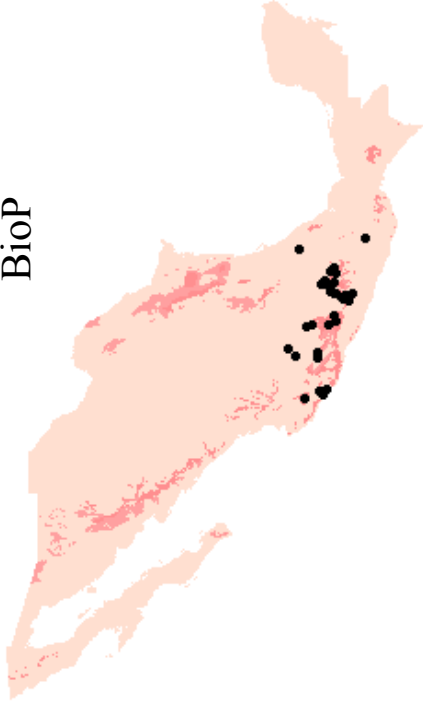


*Baronia brevicornis*: Distribuciones predichas

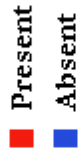
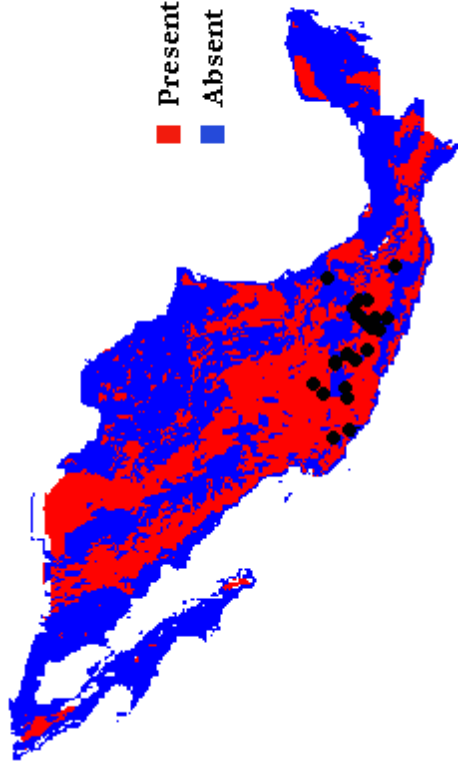
Domain



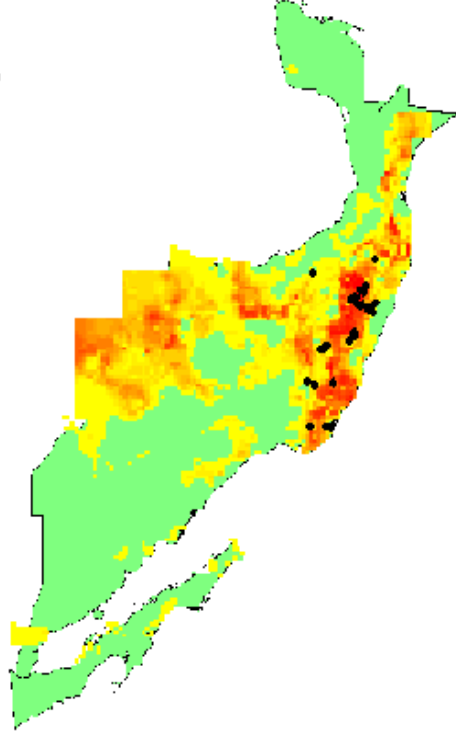
BioP



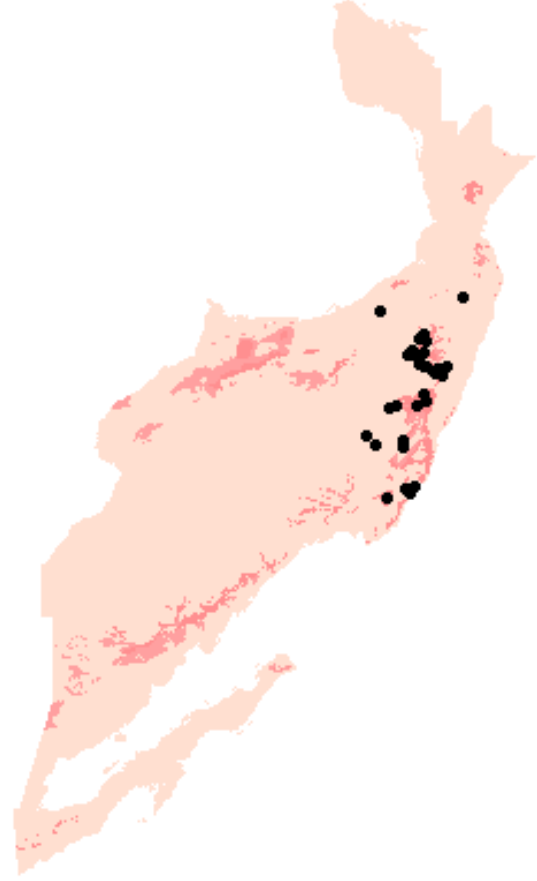
GARP



FloraMap



# Distribución predicha en detalle



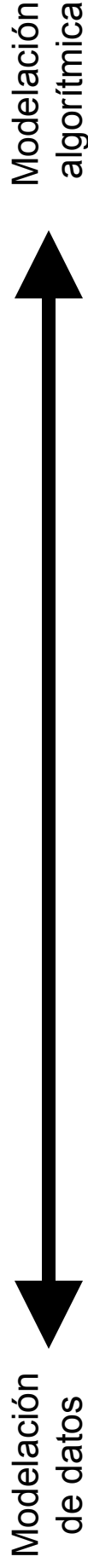
Estimación



Precisión



# Tira y afloja DM vs. AM



<b>Pros</b>	Cuantificación de incertidumbre es parte inherente del juego.	Cuantificación de incertidumbre frecuentemente faltante, difícil, o imposible.	<b>Contras</b>
	Contexto específico del problema se examina con mayor detalle, permitiendo se descubran aspectos relevantes.	Los métodos son demasiado generales, de tal manera que aspectos relevantes son ignorados.	
	Puede proveer de entendimiento estructural al problema.	Casi siempre es difícil de interpretar.	
<b>Contras</b>	No es adecuado para grandes números de casos, es decir, procesamiento en serie o experimentación.	Adecuado para procesamiento automático, en serie, sin supervisión estrecha.	<b>Pros</b>
	Puede ser entorpecido por grandes cantidades de datos o dimensionalidad.	Generalmente existe software general bien probado.	
	Primero debe meditar sobre naturaleza del modelo antes de empezar; quizás requiere de más información.	Sólo requiere de datos numéricos como entradas.	

## Resumen de conclusiones

- “algoritmos”, “datos”, y “modelación” se colocan en distintos niveles lógicos
  - En DM, el algoritmo es prescrito *ad hoc* como parte de la caja negra; en AM es la caja negra en sí.
  - AM generalmente comienza con datos; DM generalmente comienza con un contexto y una pregunta o una hipótesis científica.

# Resumen de conclusiones

- Diferentes requerimientos de “datos” según modeladores de distintas culturas de modelación
  - DM enfatiza de mayor manera el contexto y el proceso explicativo de datos, además de los valores numéricos (por qué, cómo, además de dónde y cuándo).
  - Los datos son mucho más que meros números.

# Algunas referencias

- Cox, D.R. (1990), “Role of Models in Statistical Analysis”, *Statistical Science*, 5, 169–174.
- Breiman, L. (2001), “Statistical Modeling: The Two Cultures”, *Statistical Science*, 16, 199–226.
- Friedman, J.H. (1997), “Data Mining and Statistics: What’s the Connection?”, *Department of Statistics and Stanford Linear Accelerator Center*, Stanford University.
- MacKay, R.J. and Oldford, R.W. (2000), “Scientific Method, Statistical Method, and the Speed of Light”, *Statistical Science*, 15, 224–253.
- Ripley, B.D. (1993), “Statistical Aspects of Neural Networks”, in *Networks and Chaos—Statistical and Probabilistic Aspects*, eds. O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall, Chapman and Hall, 40–123.
- Sprott, D. A. (2000), *Statistical Inference in Science*, Springer-Verlag, New York.
- Argáez, J., Christen, J.A., Nakamura, M. and Soberón, J. (2005), “Prediction of Potential Areas of Species Distributions Based on Presence-only Data”, *Journal of Environmental and Ecological Statistics*, vol. 12, 27–44.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.
- Duda, R.O., Hart, P.E., and Stork, D.G. (2001), *Pattern Classification*, Wiley, New York.
- Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.
- Stone, M. (1974), “Cross-validatory choice and assessment of statistical predictions”, *Journal of the Royal Statistical Society*, 36, 111–147.
- Breiman, L., and Spector, P. (1992), “Submodel selection and evaluation in regression: the X-random case”, *The International Statistics Review*, 60, 291–319.
- Efron, B. (1986), “How biased is the apparent error rate of a prediction rule?”, *Journal of the American Statistical Association*, 81, 461–470.