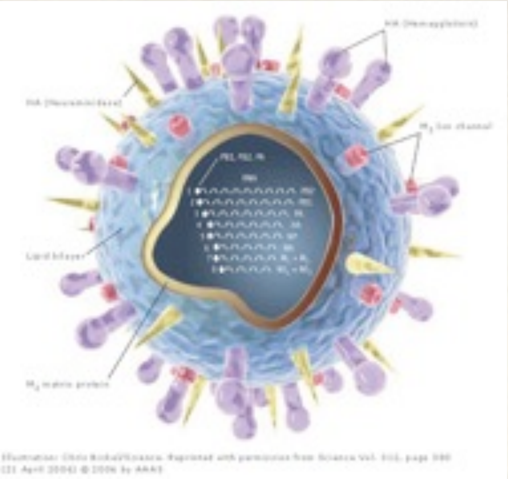
CONABIO

Spatial Modelling Using SPECIES

Sistema Para la ExploraCión de Informacion ESpacial

Chris Stephens
C3-Centro de Ciencias de la Complejidad y Instituto de Ciencias Nucleares, UNAM

Taller SPECIES - C3-CONABIO C3 13/09/2016

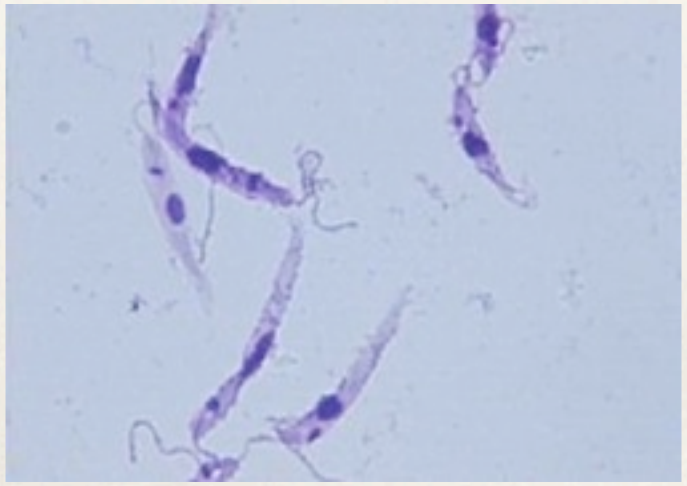


Ecology is the scientific analysis and study of <u>interactions</u> among organisms and their environment

Type of interaction	Sign	Effects
mutualism	+/+	both species benefit from interaction
commensalism	+/0	one species benefits, one unaffected
competition	-/-	each species affected negatively
predation, parasitism, herbivory	+/-	one species benefits, one is disadvantaged

Multifactorial with changing interactions

An Ecology is a Complex Adaptive System



Just how many interactions can we directly observe?

Importancia médica

T. Infestans

T. barberi

T. pallidipennis

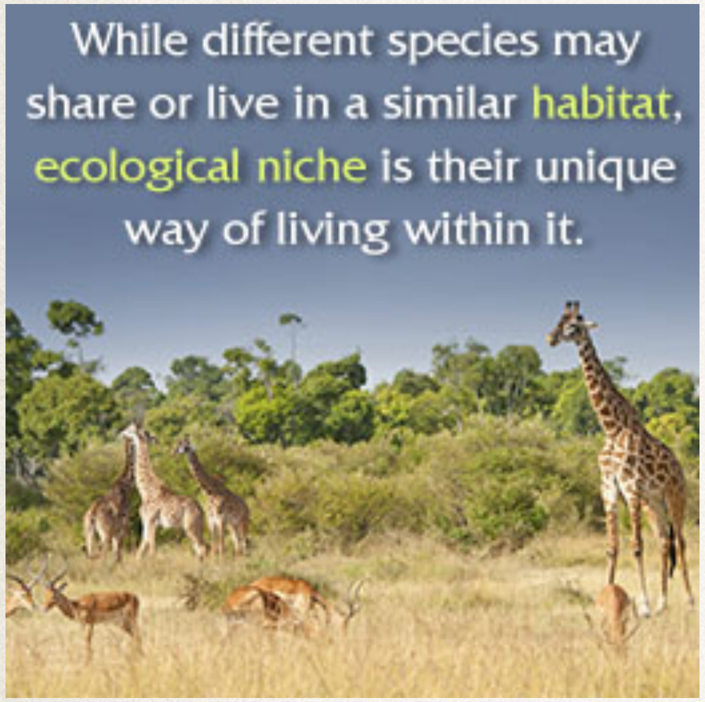
T. longipenis

T. recurva

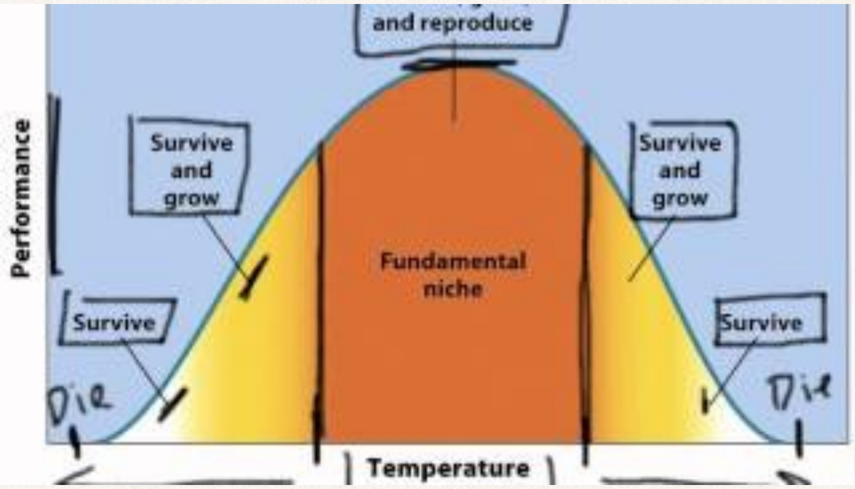
T. neotomac

Culex quinquefasciatus

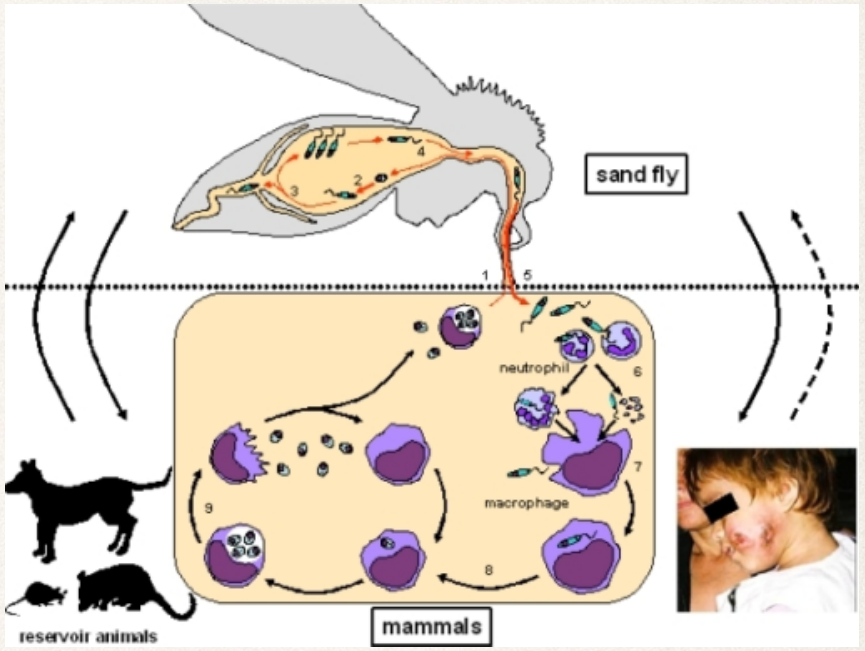
Niche versus Community



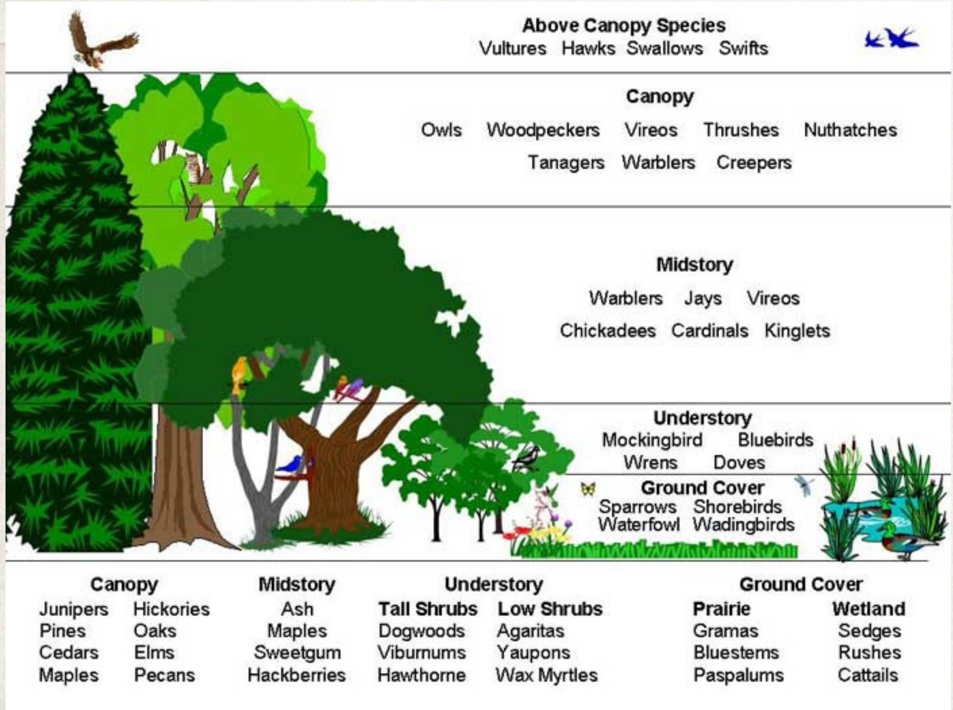
While different species may share or live in a similar habitat, ecological niche is their unique way of living within it.



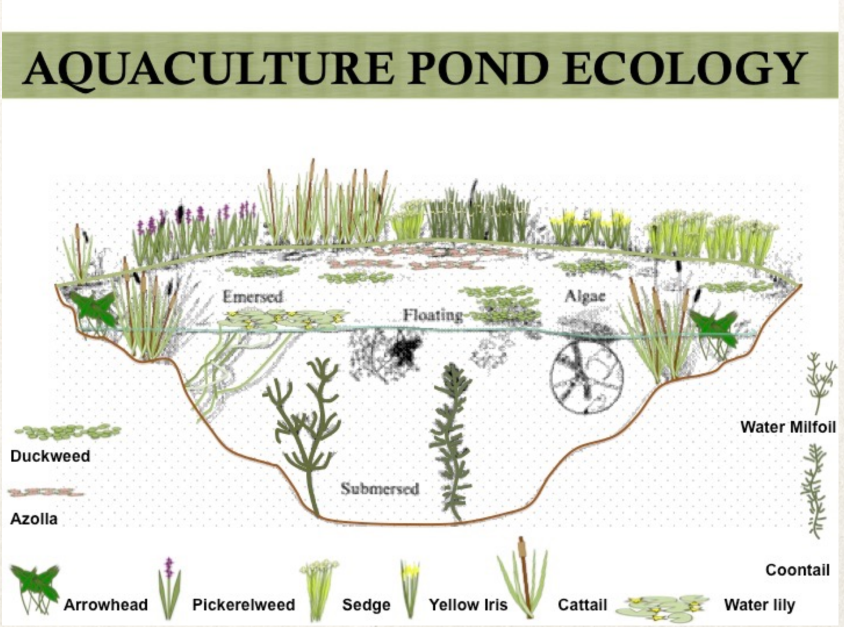
Hutchinson: "the set of biotic and abiotic conditions in which a species is able to persist and maintain stable population sizes."



Community ecology examines how interactions among species and their environment affect the abundance, distribution and diversity of species within communities.



Canopy		Midstory	Understory		Ground Cover	
			Tall Shrubs	Low Shrubs	Prairie	Wetland
Junipers	Hickories	Ash	Dogwoods	Agaritas	Gramas	Sedges
Pines	Oaks	Maples	Viburnums	Yaupons	Bluestems	Rushes
Cedars	Elms	Sweetgum	Hawthorne	Wax Myrtles	Paspalums	Cattails
Maples	Pecans	Hackberries				



Community Ecology

- A *community* is an assemblage of species (populations) living close enough together for potential interaction in a habitat

Ecology is the scientific analysis and study of

interactions

among organisms and their environment

Physics is the scientific analysis and study of

interactions

between matter and energy

How have we understood **interactions** in physics?

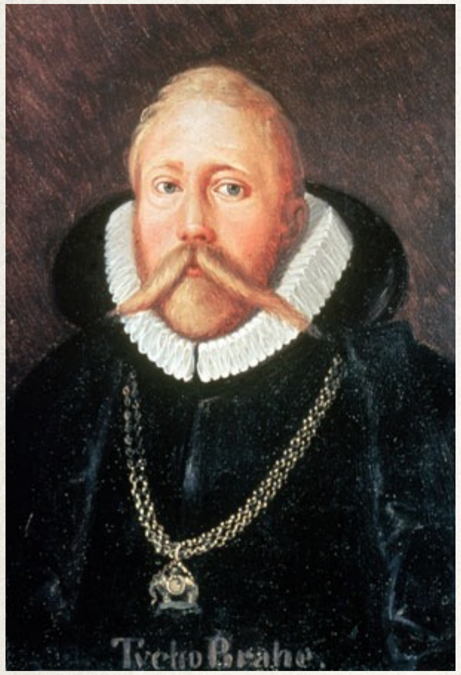
Through Spatial Modeling!

Studying where things are, and when,

relative to each other.

Spatial Modeling in the past…

Data —> Phenomenology —> Taxonomy —> Theory



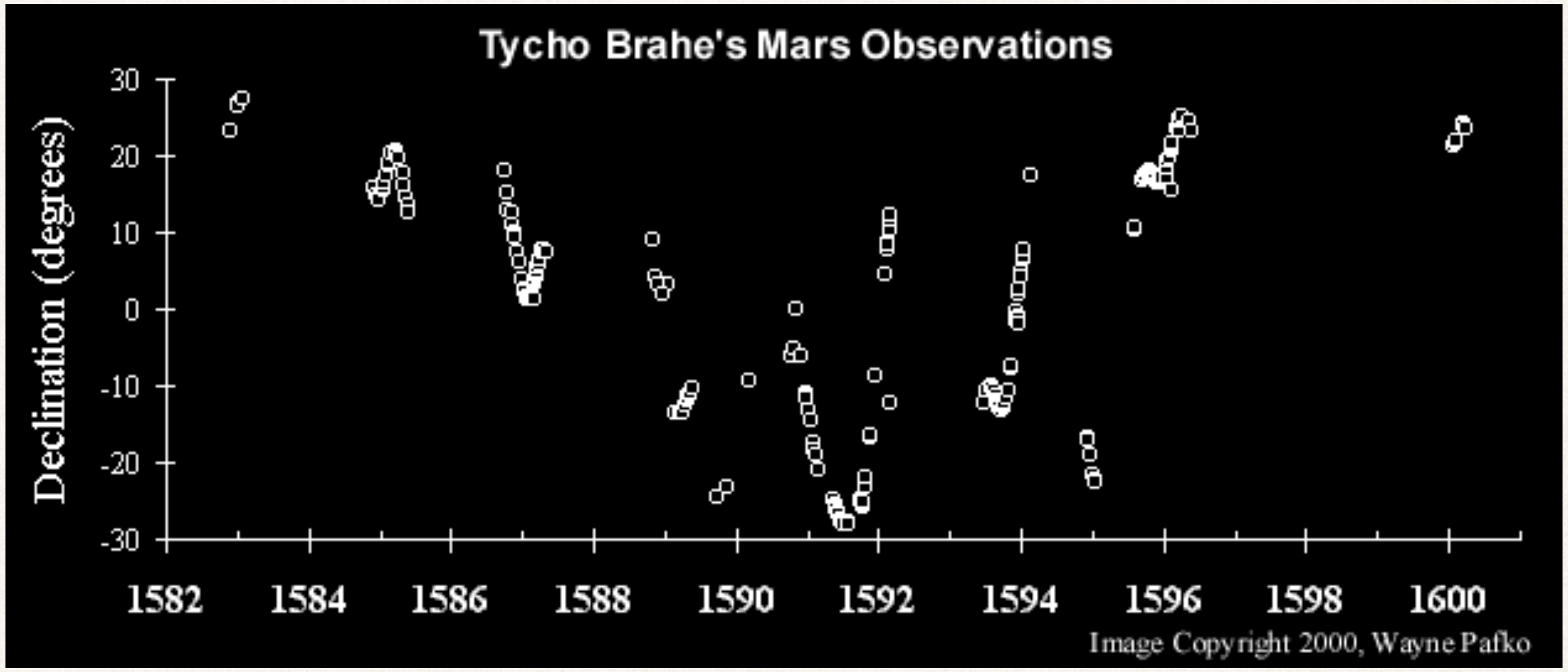
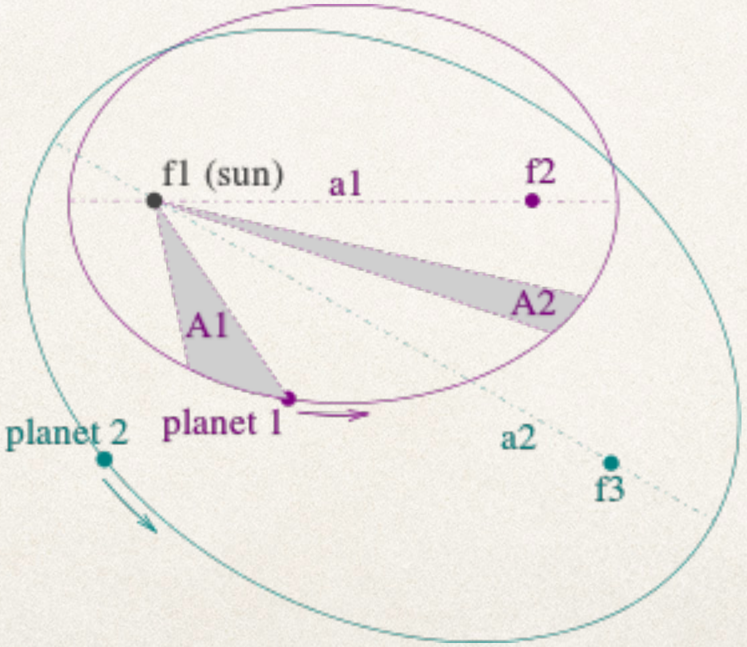


Image Copyright 2000, Wayne Pafko

Data

Phenomenology

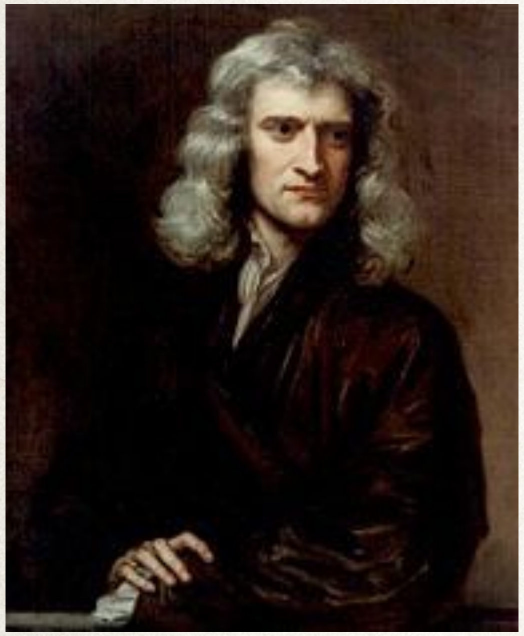


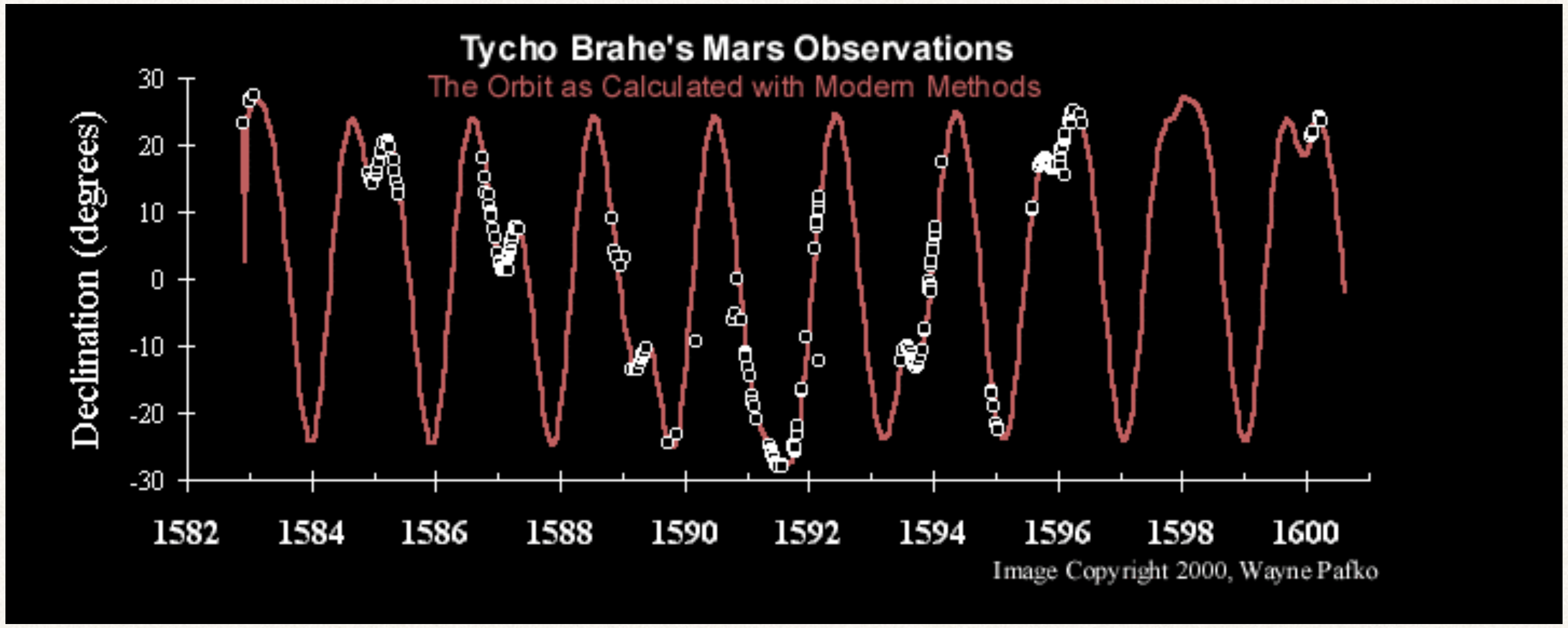
Kepler's Laws

1. The orbit of a planet is an ellipse with the Sun at one of the two foci.
2. A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.
3. The square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit.

Spatial Modeling in the past…

Data —> Phenomenology —> Taxonomy —> Theory




Tycho Brahe's Mars Observations

The Orbit as Calculated with Modern Methods

Image Copyright 2000, Wayne Pafko

Theory

$$F = ma$$

$$F = GMm / r^2$$

Isaac Newton computed the acceleration of a planet moving according to Kepler's first and second law.

1. The *direction* of the acceleration is towards the Sun.
2. The *magnitude* of the acceleration is inversely proportional to the square of the planet's distance from the Sun (the *inverse square law*).

This implies that the Sun may be the physical cause of the acceleration of planets.

Newton defined the force acting on a planet to be the product of its mass and the acceleration. So:

1. Every planet is attracted towards the Sun.
2. The force acting on a planet is in direct proportion to the mass of the planet and in inverse proportion to the square of its distance from the Sun.

The Sun plays an unsymmetrical part, which is unjustified. So he assumed, in Newton's law of universal gravitation:

1. All bodies in the solar system attract one another.
2. The force between two bodies is in direct proportion to the product of their masses and in inverse proportion to the square of the distance between them.

As the planets have small masses compared to the Sun, the orbits conform approximately to Kepler's laws. Newton's model fits actual observations more accurately.

The Difference between Physical and Complex Adaptive Systems

In Complex Adaptive Systems…

s a lot

Imagine what you can say about a city

versus

a crystal as big as a city!

Multifactoriality

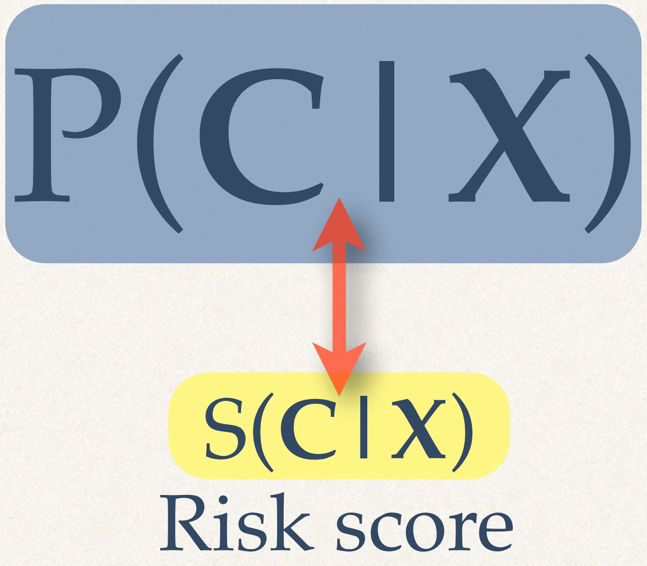
Adaptation

To say a lot, you need to have a lot of data…
Big Data… A Data Revolution!

The data revolution and the access to big, deep data is revolutionising our ability to study the immensely rich phenomenology of complex systems and construct more appropriate taxonomies

"Keplerian" Ecological models

What do we want to predict?
C = (C1, C2, C3, …, CN)
the presence, or abundance, or,… of one or more populations or taxa



$P(C|\mathbf{X})$

$S(C|\mathbf{X})$

Risk score

What affects it?
The "niche"
$\mathbf{X}$ = (X1, X2, X3, …, XM)

A large part of the complexity is in the multi-factoriality of both C and X. Adaptation is inherent in the fact that P(C|X) can change in time.

$$\mathbf{X} = X(sd)+X(se)+X(n)+X(ev)+X(g)+X(af)+X(hm)+X(i)+X(sp)+\ldots$$

Macro-Climactic factors

Micro-Climatic factors

Hydrography

Prey species

Human activity

Behavioural characteristics

Phenotypic characteristics

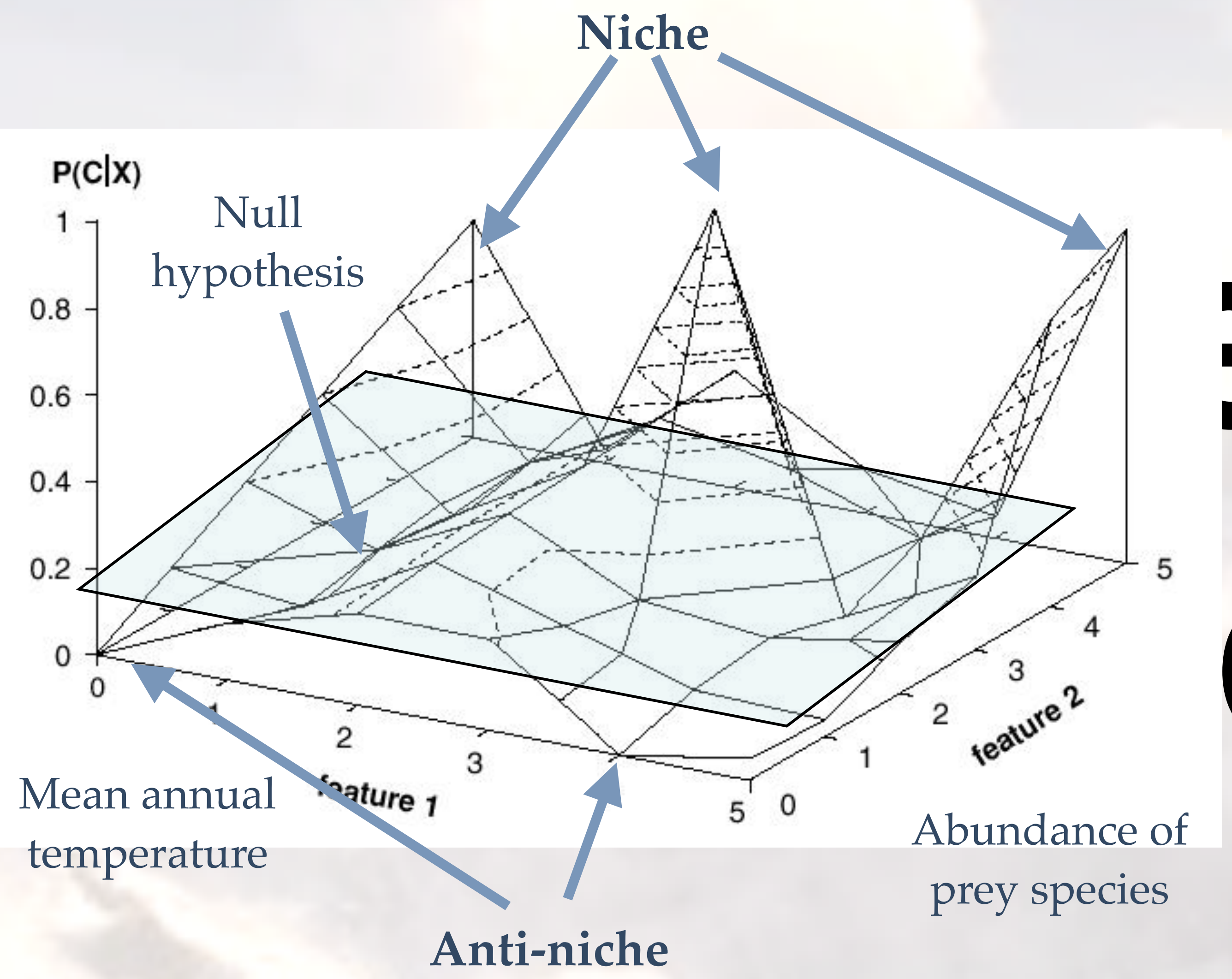
Competitor species

Predator species

Problems of co-dependence and causality

Are there generic topologies for Niche or Ecosystemic landscapes?

Can they be multi-modal?



Niche

Null hypothesis

Mean annual temperature

Abundance of prey species

Anti-niche

Are they rugged or smooth?

What are the “right” coordinates?

What are the patterns of epistasis?

And the data? Where are the “Brahes”? There’s lots of them!

Normally data mining takes place in a “categorical” space (the equivalent in ecology is a niche space). However, most ecological data is spatio-temporal at multiple scales. Spatial data mining is much less developed than standard data mining.

- Collection data ← SNIB, CONABIO
- Ecological niche data
- Ecological niche model data
- Socio-economic data
- Socio-demographic data
- Phenotypic data
- Vegetable and crop cover
- Geographical data
- Medical and public health data…

Problems with spatial data:

Different sources

Different location, data base, access,…

Different data types

categorical, metric, continuous, discrete,…

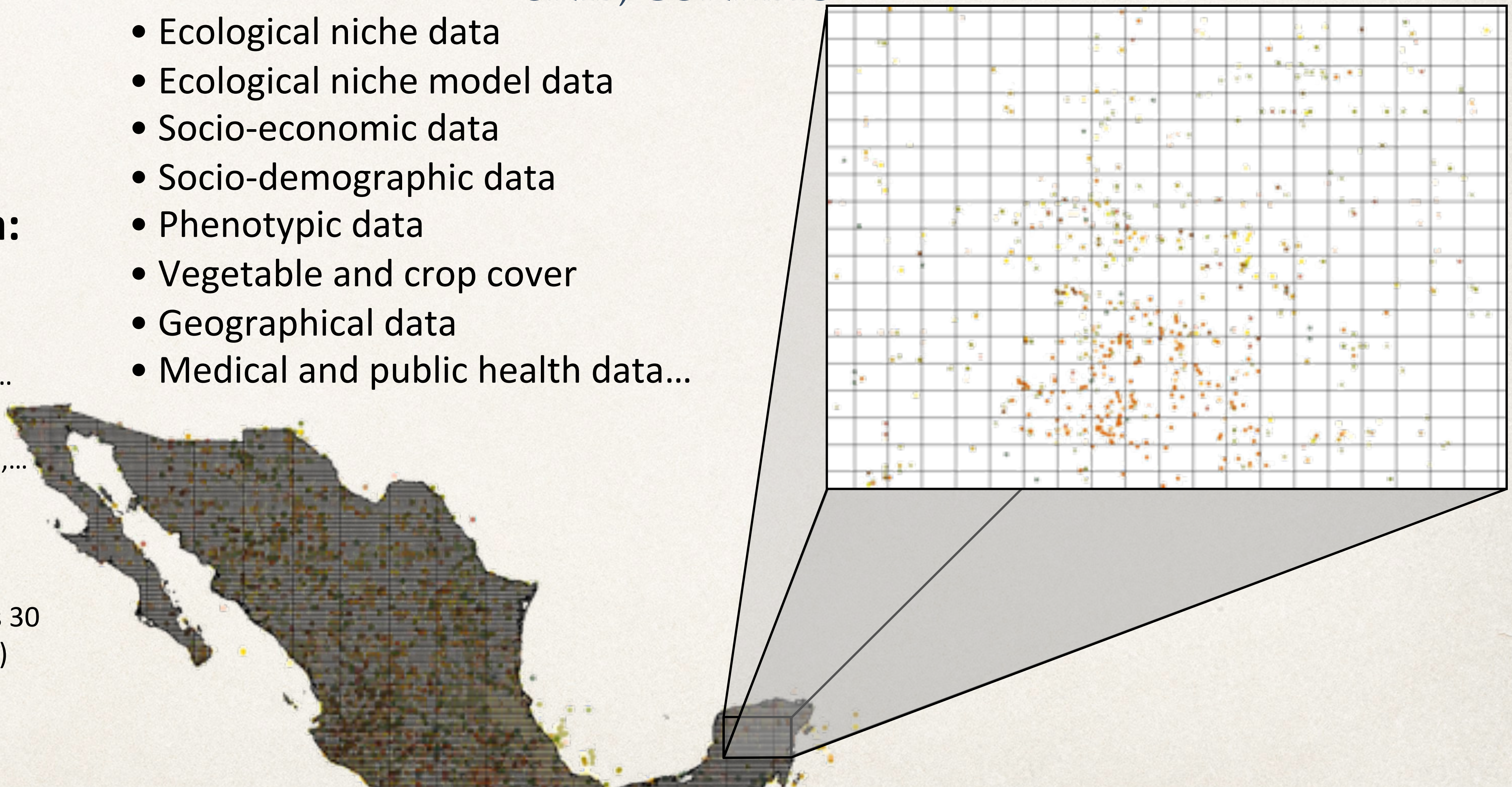
Different spatial resolution

Explicit – e.g., pixel by pixel in environmental layers

Implicit – 30,000,000 data points versus 30

“Quality” (e.g. Phenotypic characteristic) versus “quantity”

Abiotic versus biotic

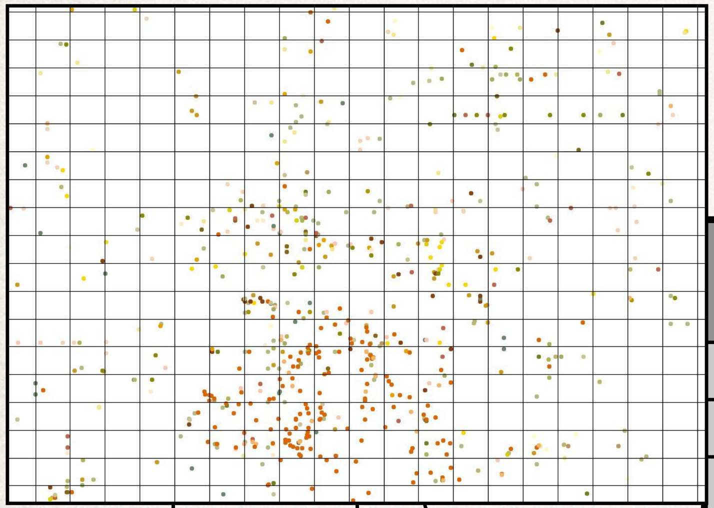


The data are represented in space and time – spatial data mining

CONABIO

A Democracy of the Data:

To infer interactions from where "things" are



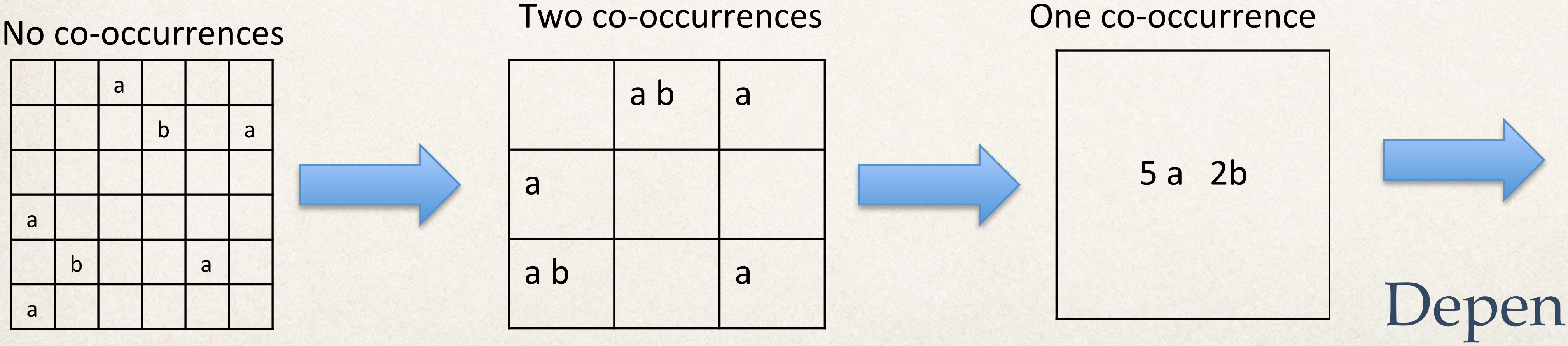
Choose a spatial resolution: give everyone one vote there.
The "Senate" versus the "Congress" approach!

Cuadrante	Sigmodon hispidus	Dipetalogaser maxima	Casos Chagas	Precipitación anual	Temperatura promedio	GARP Triatoma maximus	GARP Diptaloster maxima	Perfil agricola
A1	1	3	1	23	18.6	1	1	4
A2	0	1	0	23	18.6	1	1	4
A3	0	2	0	23.7	18.7	1	1	1
A4	0	4	0	23.7	18.7	1	1	3
A5	0	2	1	23.7	18.7	1	1	3
A6	2	5	2	23.7	18.7	1	1	2
A7	0	1	0	23.3	18.4	1	1	5
A8	0	2	0	22.8	18.8	1	1	3
A9	1	3	1	22.8	18.8	1	1	1
A10	0	1	0	22.8	18.8	0	1	1
A11	0	0	0	22.8	18.8	0	1	1
A12	0	0	0	22.8	18.8	0	1	2
A13	0	0	0	22.8	18.8	0	0	4
A14	0	0	0	22.8	18.8	0	0	3
A15	0	2	0	22.8	18.8	0	1	4
A16	0	1	0	22.8	18.8	0	1	2
A17	0	0	0	22.8	18.8	0	1	1
A18	0	0	0	22.8	18.8	0	0	1

Now we can make statistical inferences

In standard data mining, for example: P(death | age) = N(death,age)/N(age); P(death | diabetes); P(death | age,diabetes); to **infer** that age is a risk factor for death, as is diabetes. Here, we count individuals who have different traits. There is a preferred statistical unit - the individual within which we can look for coincidences/co-occurrences. In spatial data mining this is not the case.

We must define coincidences/co-occurrences using an appropriate **uniform** spatio-temporal scale.



No co-occurrences

		a			
			b		a
a					
	b			a	
a					

Two co-occurrences

	a b	a
a		
a b		a

One co-occurrence

5 a 2b

N(ab)/N(b) = P(a | b)

Dependence of species a on niche variable b

Here we're in geographic space

The Technical Part

How do we decide if the frequency of co-occurrence $P(a|b)$ is less or more than “expected?

Its just like flipping a coin! A binomial process. How many times when I flip a coin of “type b” do I get result “a”?

What’s my baseline, my expectation, my “null hypothesis”?

That b does not “influence" a, so $P(a|b) = P(a)$. So, is $(P(a|b) - P(a))$ “big”?

$$\mathrm{epsilon}(a|b) = N(b)(P(a|b) - P(a))/(N(b)P(a)(1-P(a)))^{1/2}$$

Standard deviation of binomial distribution. The right unit to measure big versus small.

If $|\mathrm{epsilon}(a|b)| > 1.96$, with 95% confidence we can reject the null hypothesis —> possible “interaction” between a and b

CONABIO

The Technical Part

But what about $P(\mathbf{C}|\mathbf{X}) = P(C|X1,X2,X3,\ldots,XN)$

$= N(CX1,X2,X3,\ldots,XN)/N(X1,X2,X3,\ldots,XN)$

Well… $N(CX1,X2,X3,\ldots,XN) = 0, 1$

the "curse of dimensionality"

→ Use Bayes' theorem

$$P(\mathbf{C}|\mathbf{X}) = P(\mathbf{X}|C)P(C)/P(\mathbf{X})$$

and assume $P_{NB}(\mathbf{X}|C) = \prod_{i=1}^{N} P(X_i|C)$

Naive Bayes Approximation

Total factorisation

$$S(C|\mathbf{X}) = \ln(P(C|X)/P(\underline{C}|X)) = \ln(P(\mathbf{X}|C)P(C)/P(\mathbf{X}|\underline{C})P(\underline{C}))$$

$$= \sum_i \ln(P(X_i|C)/P(X_i|\underline{C})) + \ln(P(C)/P(\underline{C}))$$

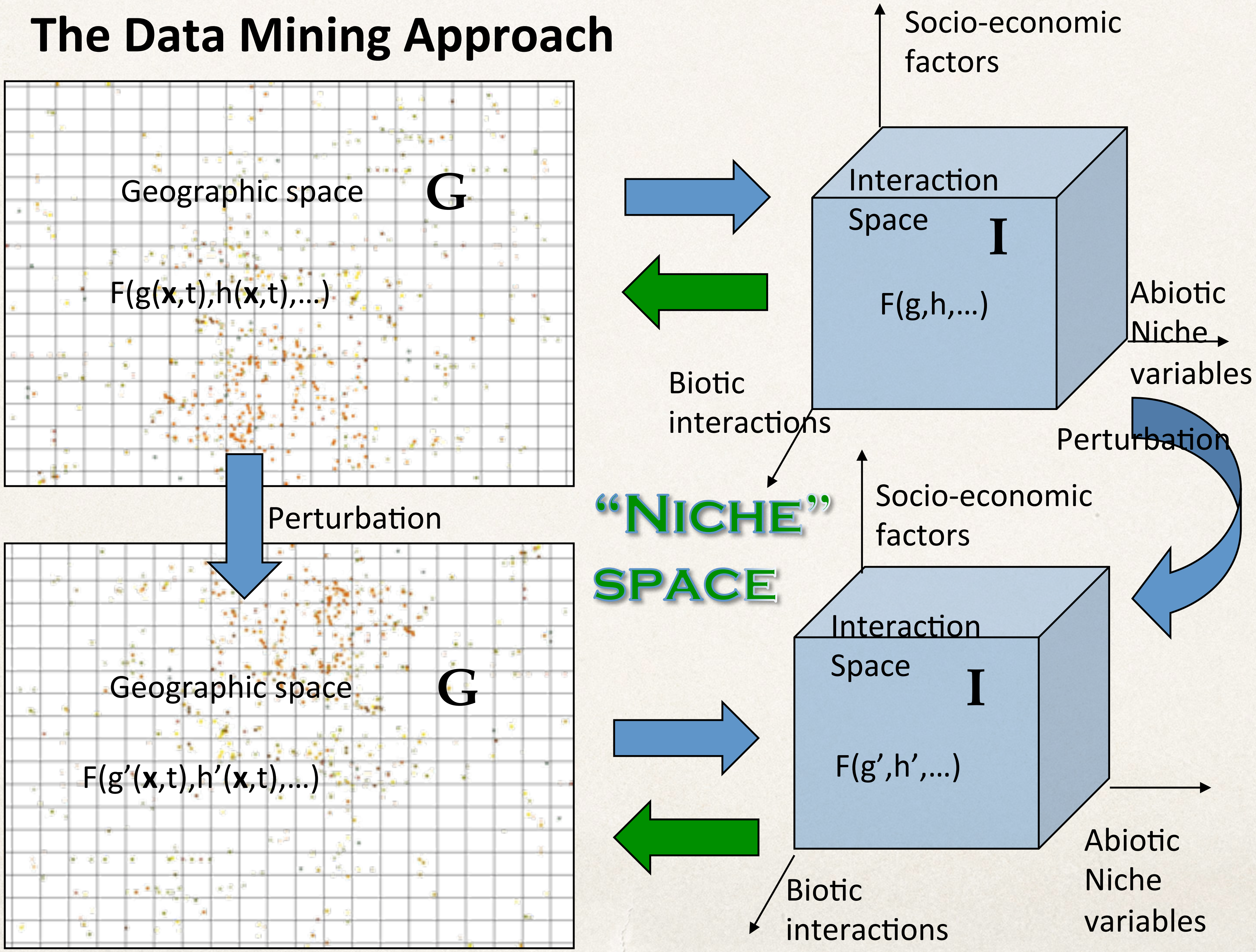
$$= \sum_i S(C|X_i) + \ln(P(C)/P(\underline{C}))$$

contribution to probability to find C from presence of niche variable X_i . Can compare contributions from biotic/abiotic/topographic/… factors

Here we're in niche space

So we can pass from Geographic space to Niche Space and vice versa

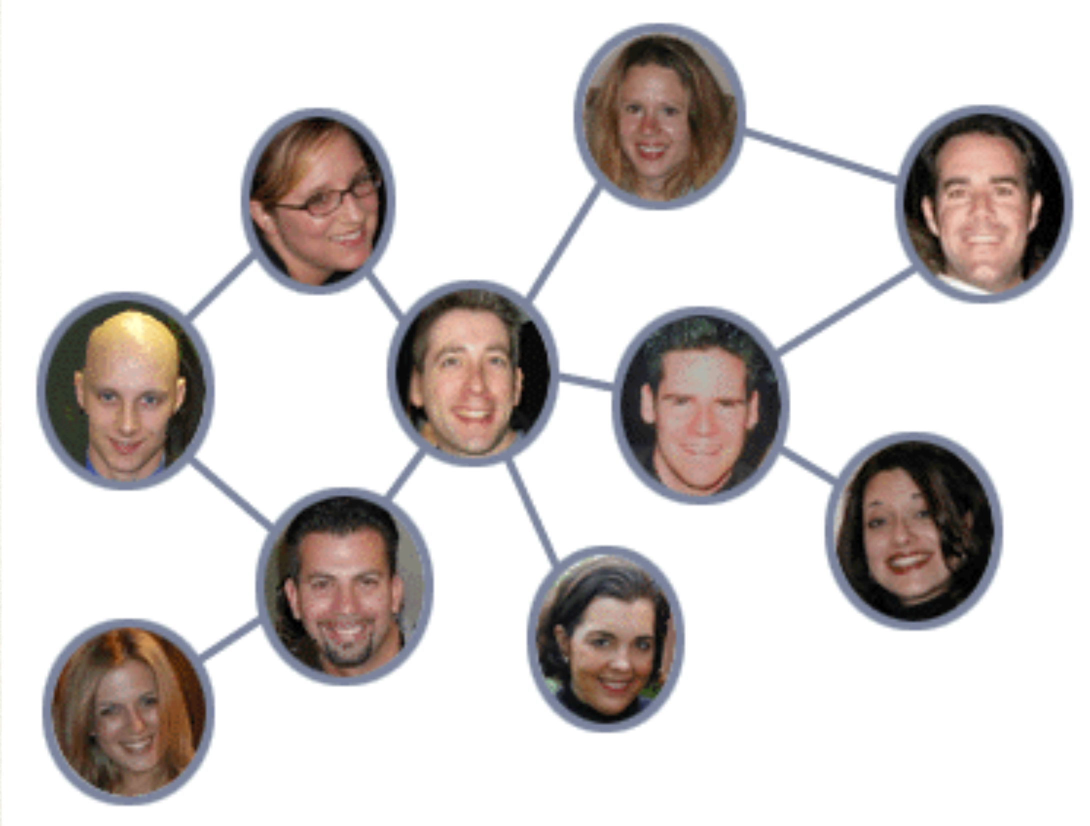
The Data Mining Approach



Now for Communities…

You can judge a man by his “friends”

or his “enemies”, or “parasites”, or “prey” or “predators” or…

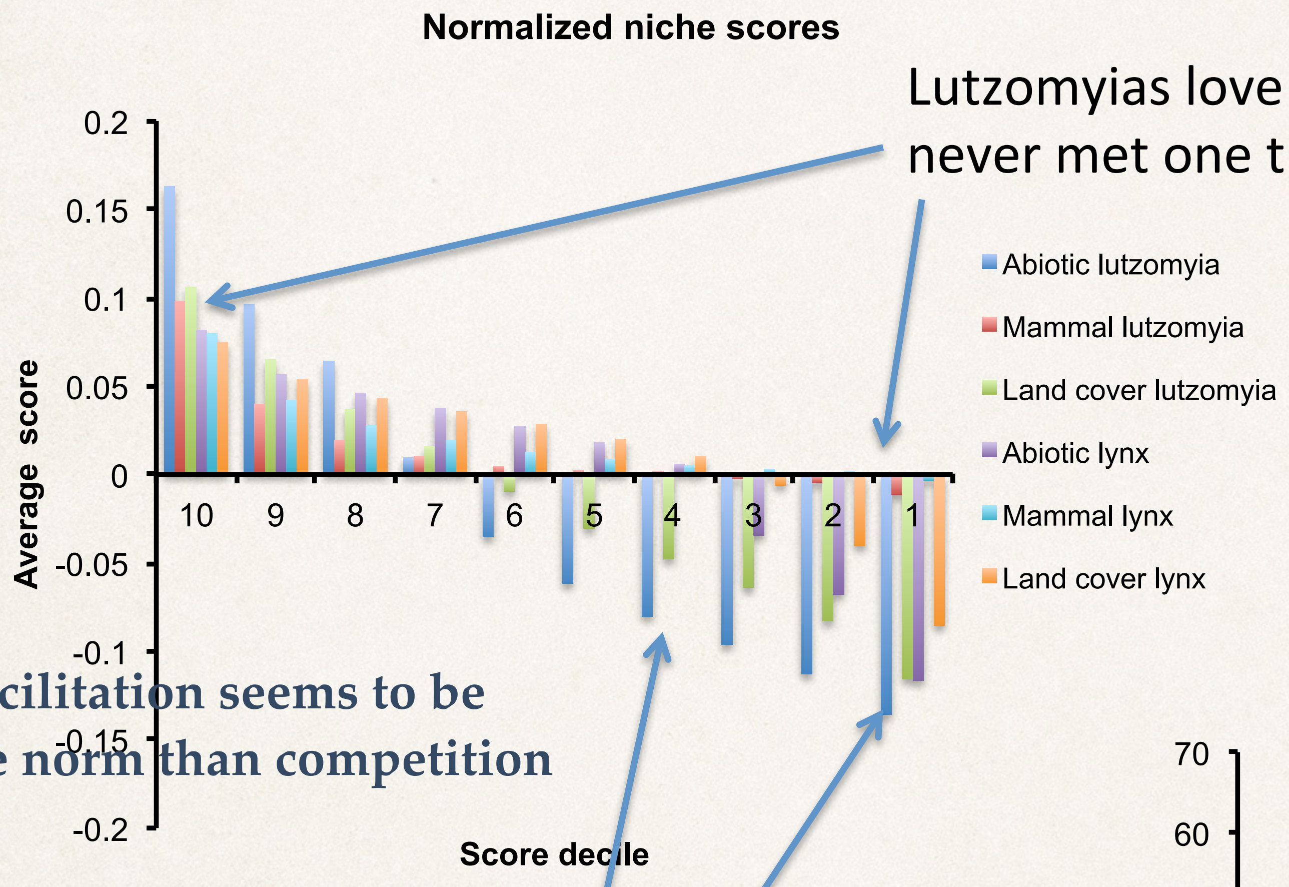


CONABIO

Use Complex Inferential Networks to Represent Community Interactions

- Take nodes to be…
  - Species, other taxonomic or phylogenetic groupings, groupings by phenotypic characteristics,
- Take links to be a statistical measure of spatial (temporal) co-occurrence
  - P(Y|X), epsilon(Y|X), P(A,B|C,D), epsilon(Z|X,Y)
  - What is a high/low degree of co-occurrence? (Choosing a null hypothesis)
  - What spatial (temporal) resolution? (When do things co-occur?)

Two Example Niches



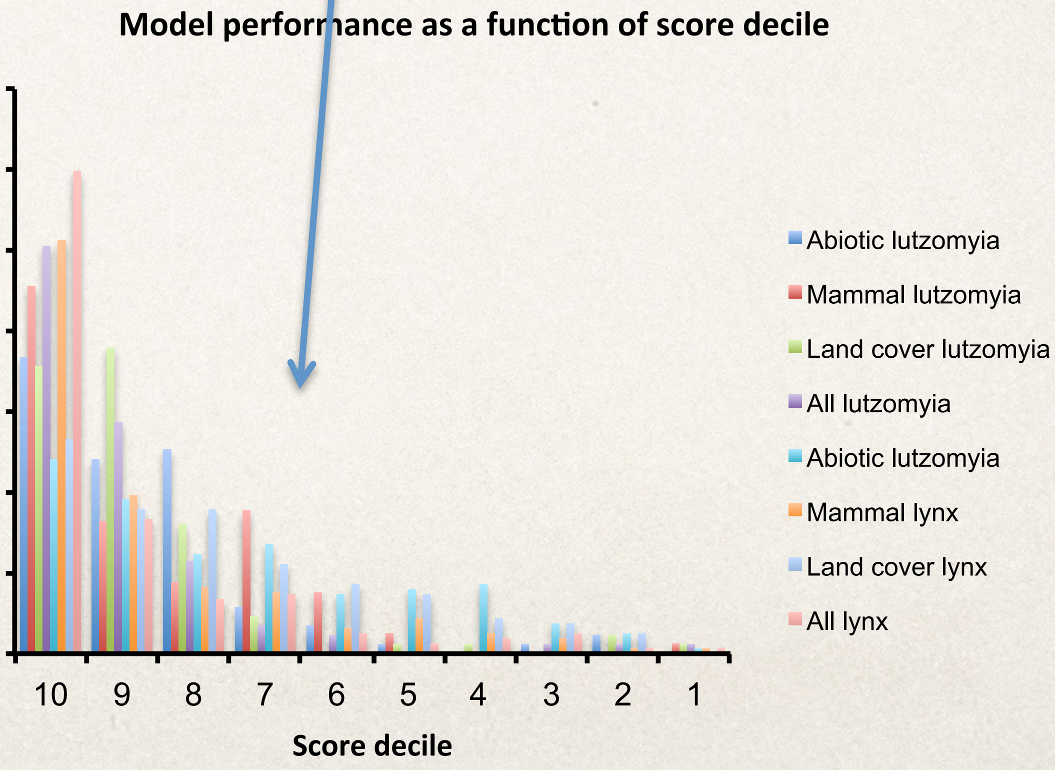
Normalized niche scores

Lutzomyias love mammals, never met one they didn't like

Biotic facilitation seems to be more the norm than competition

Climatic factors are more important for determining where Lutzomyias aren't rather than where they are

Including in a fuller, richer Niche Space leads to more predictive models (less false positives/negatives)

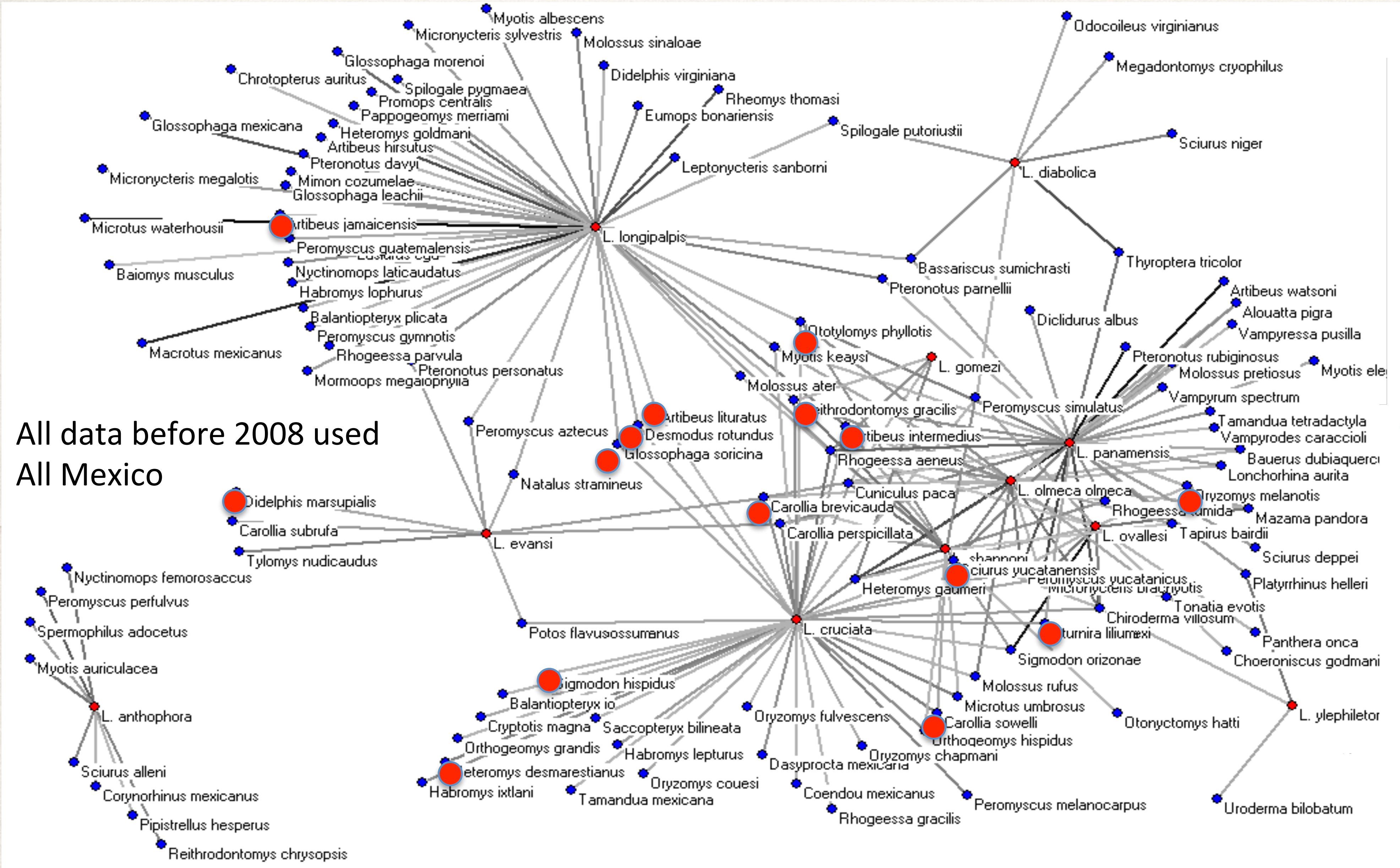


Model performance as a function of score decile

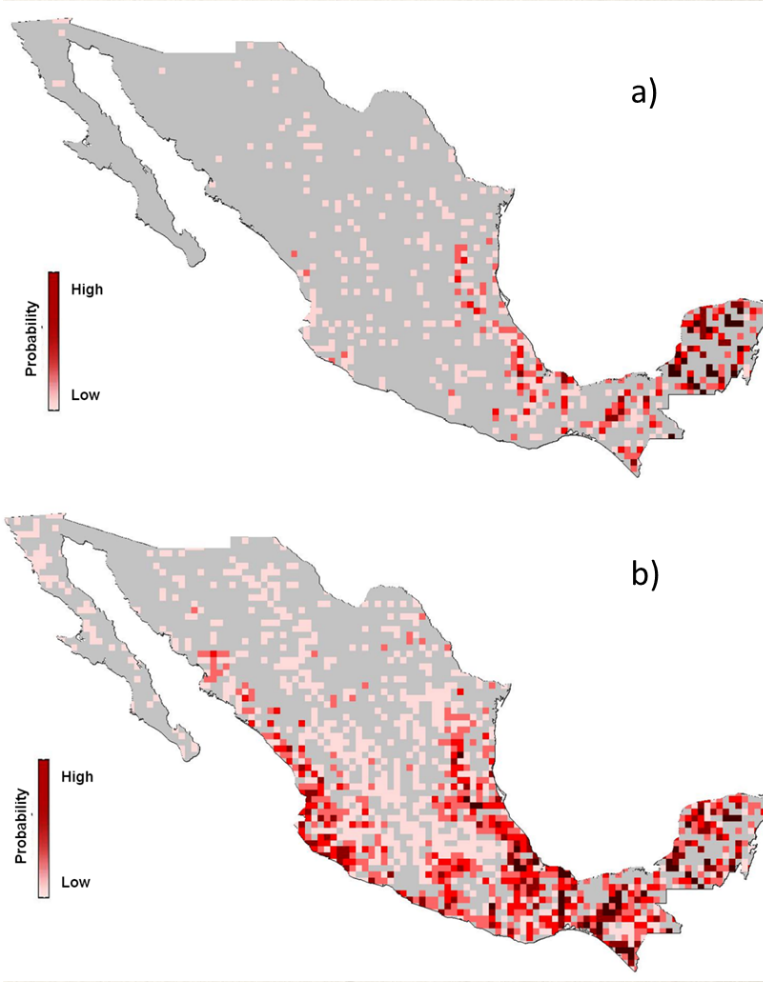
Chains of causality

CONABIO

The Ecology of Leishmaniasis



All data before 2008 used
All Mexico



What does this tell us about vector control?

Conclusions: CAS

- **All science is Data Science!**
  - The difference now is the big, deep data available due to the Data Revolution
  - Much of this data is spatio-temporal - where "things" are and when
  - Data associated with the relative positions of "things" in space and time has allowed us to deduce (Data —> Phenomenology —> Taxonomy —> Theory) the nature of the interactions between physical objects: the four fundamental forces
  - These forces are universal and simple
- **Unlike the physical world, ecologies are CAS composed of other CAS**
  - We don't have adequate conceptual or theoretical frameworks in which to understand CAS
  - The phenomenology of CAS is incredibly rich and qualitatively different from that of physical systems (multi-factorial from the micro to the macro, and adaptive)
  - To describe this phenomenology you need a lot of data

Conclusions: Ecology

- Spatio-temporal data about organisms, relative to each other (biotic) and relative to the environment (abiotic), can be used to deduce the nature of the interactions between them and with the environment
  - This can be done at the niche level (one to many) and at the community level (many to many)
  - Our formalism allows for the incorporation of any data type, data format and data resolution
- The Niche "fitness" landscape of a taxon C can be characterised quantitatively by $P(C|\mathbf{X})$ using spatio-temporal data mining
  - What are their general topological and geometrical characterisations?
  - How rugged/smooth are they?
  - What is the distribution of epistasis
    - Are distributions random?
    - Facilitation versus competition
  - What are the right coordinates?
  - What is the dynamics of Niche landscapes? How do they evolve?
  - How do we determine and characterise causal chains in ecology?

Conclusions: Ecology

- At the community level, spatio-temporal data can be used to construct **Complex Inference Networks** (CIN) as representations of communities and ecosystems
    - How to distinguish causality from correlation?
    - How to determine co-dependencies?
- As a proof of concept: The niches and community relations of diseases can be determined via CIN
    - Identification of transmission cycles and host range
        - Leishmania, Chagas, Lyme, Dengue, Zika, West Nile,…
    - Many zoonoses are multi-host, multi-vector, multi-pathogen systems.

CONABIO

Grupo de Trabajo

C3 - Centro de Ciencias de la Complejidad, UNAM; Instituto de Biología, UNAM; CONABIO; Universidad Catolica de Chile; Facultad de Medicina, UNAM

1.- Dr. Christopher R. Stephens
2.- Dr. Raúl Sierra Alcocer
4.- Dr. Constantino González Salazar
5.- M. en C. Enrique del Callejo
6.- M. en C. Everardo Robredo
7.- Lic. Juan Carlos Salazar Carrillo

Publications

Competitive interactions between felid species may limit the southern distribution of bobcats Lynx rufus
V Sánchez-Cordero, D Stockwell, S Sarkar, H Liu, CR Stephens, ...
Ecography 31 (6), 757-764, 2008

Using biotic interaction networks for prediction in biodiversity and emerging diseases
CR Stephens, JG Heau, C González, CN Ibarra-Cerdeña, ...
PLoS One 4 (5), e5725, 2009

Exploratory analysis of the interrelations between co-located boolean spatial features using network graphs
R Sierra, CR Stephens
International Journal of Geographical Information Science 26 (3), 441-468, 2012

Constructing ecological networks: a tool to infer risk of transmission and dispersal of Leishmaniasis
C González-Salazar, CR Stephens
Zoonoses and public health 59 (s2), 179-193, 2012

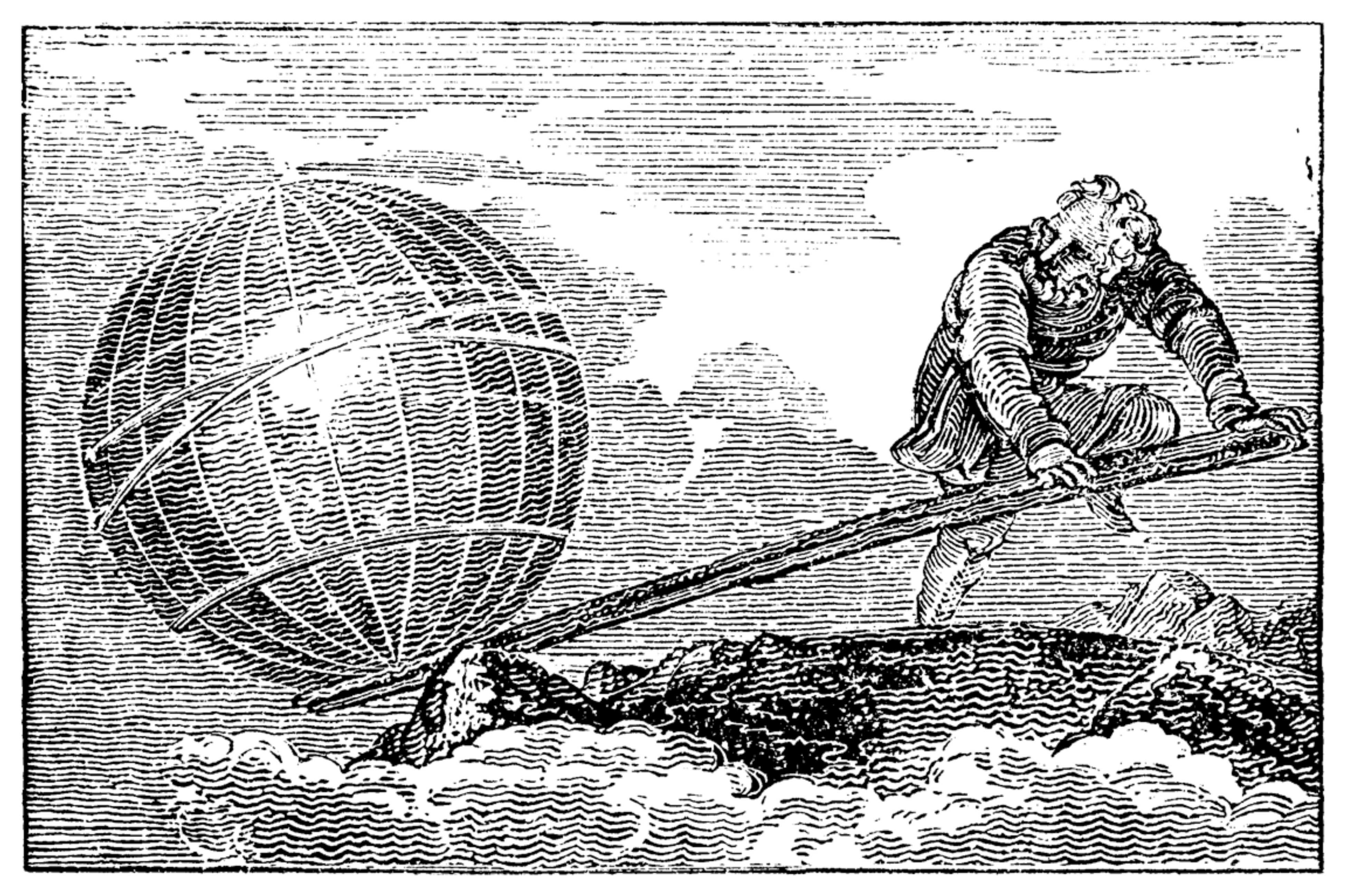
Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions
C González-Salazar, CR Stephens, PA Marquet
Ecological Modelling 248, 57-70, 2013

Leishmania (L.) mexicana Infected Bats in Mexico: Novel Potential Reservoirs
M Berzunza-Cruz, Á Rodríguez-Moreno, G Gutiérrez-Granados, ...
PLoS neglected tropical diseases 9 (1), e0003438-e0003438, 2015

Predicting the potential role of non-human hosts in Zika virus maintenance
C González-Salazar, CR Stephens and V. Sanchez-Cordero
submitted to Eco-health

UNDERSTANDING TRANSMISSIBILITY PATTERNS OF CHAGAS DISEASE THROUGH COMPLEX VECTOR-HOST NETWORKS
Laura Rengifo-Correa, Constantino González-Salazar, Juan J. Morrone, Juan Luis Téllez-Rendón, Christopher Stephens, submitted to PLoS Neglected Tropical diseases

Can you judge a disease host by the company it keeps? Predicting disease hosts and their relative importance using complex networks
CR Stephens et al, submitted to PLoS Neglected Tropical diseases



δῶς μοι πᾶ στῶ καὶ τὰν γᾶν κινάσω

Give me a place to stand on and I´ll move the earth

Give me enough data and I´ll predict anything

The Data Revolution will revolutionise our ability to model and understand ecology