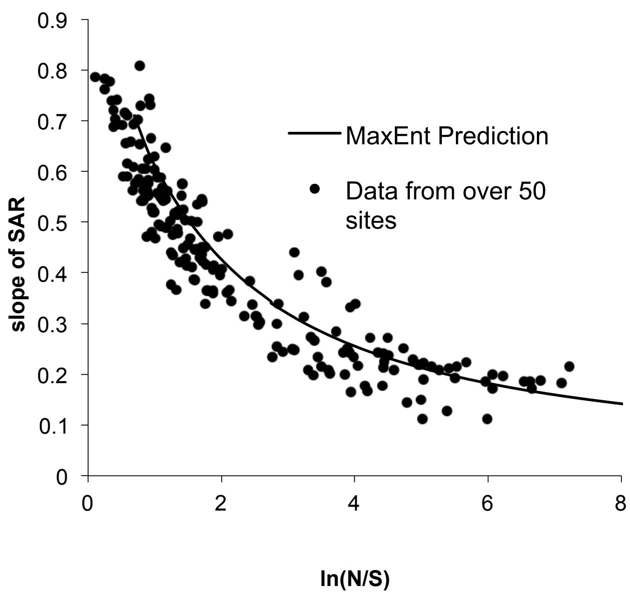
On the Inference of Pattern and Process in Nature:

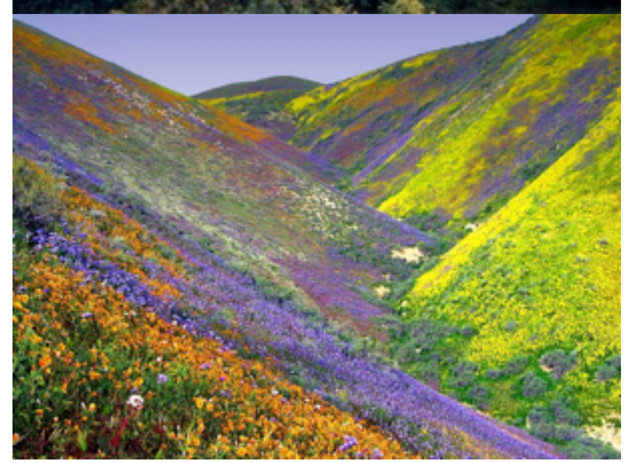
What Information Theory Can Teach Us

John Harte

CSSS, Chile

November 12, 2013





Outline of CSSS Lectures on MaxEnt and Macroecology

PART I

1. **Introduction:** Ecological Complexity

2. **The metrics of macroecology**: definitions, prevailing patterns

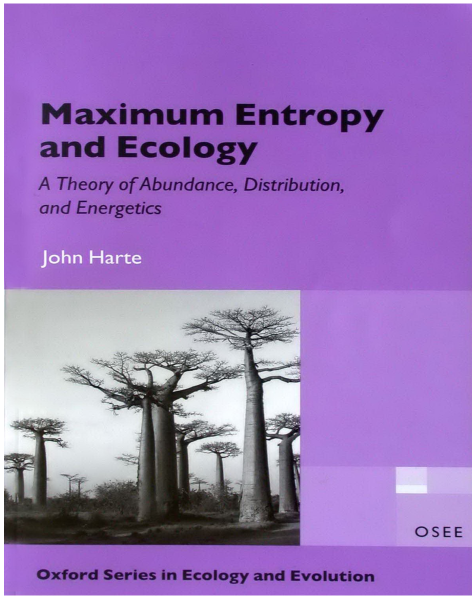
3: **MaxEnt**: history, rationale, current applications, techniques.

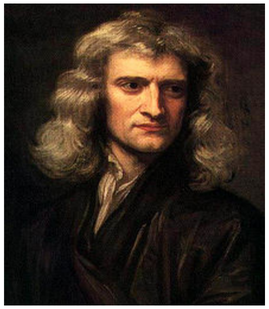
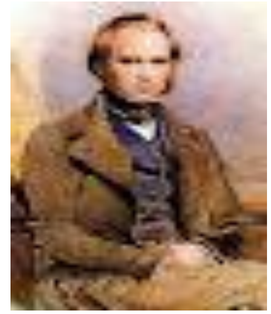
PART II

4: **MaxEnt Theory of macroecology:** structure of the theory, predictions.

5: **MaxEnt theory of macroecology:** tests of theory

6. **At the frontier.**



PHYSICS	ECOLOGY
The more you look, the simpler it gets	The more you look, the more complex it gets
Primacy of initial conditions	Primacy of contingency and history
Universal patterns; Search for mathematical laws	Weak trends; Reluctance to seek quantitative laws
Mostly predictive	Mostly explanatory
Central role for idealized systems	Reluctance to caricature nature

Why Does Ecology Appear Resistant to Theory?

- Feedback, nonlinear synergies, thresholds, and irreversibilities; a wealth of fascinating detail
- Conducting large-scale experiments is impossible
- History and Contingency; initial conditions are not enough
- Drawing space-time boundaries is difficult
- Local to Planetary scale disruption; degradation and extinction of the objects of study

The Dilemma faced by Ecosystem Modelers:

- **Many mechanisms and processes:**

  predation, mutualism, competition, dispersal, speciation,birth, death, pollination, cannibalism, migration, …

- **Many traits and behaviors:**

  body size, speed, phenology, food preferences, rooting depth, mating strategies, coloration, temperature tolerance, nutrient acquisition strategies, …

- **Stochastic environments, historical contingency**

all influence Patterns in Macroecology.

Hence basing models on explicit mechanisms, traits & behaviors generally results in

The necessity of somewhat arbitrary choices

Adjustable parameters

Models that are not readily falsifiable.

The Goal of this Work

To predict patterns in "macroecology"

- **Across taxa:** plants, bugs, birds,…
- **Across spatial scale:** small patches to large biomes
- **Across habitats:** forests, meadows, deserts, tundra,…

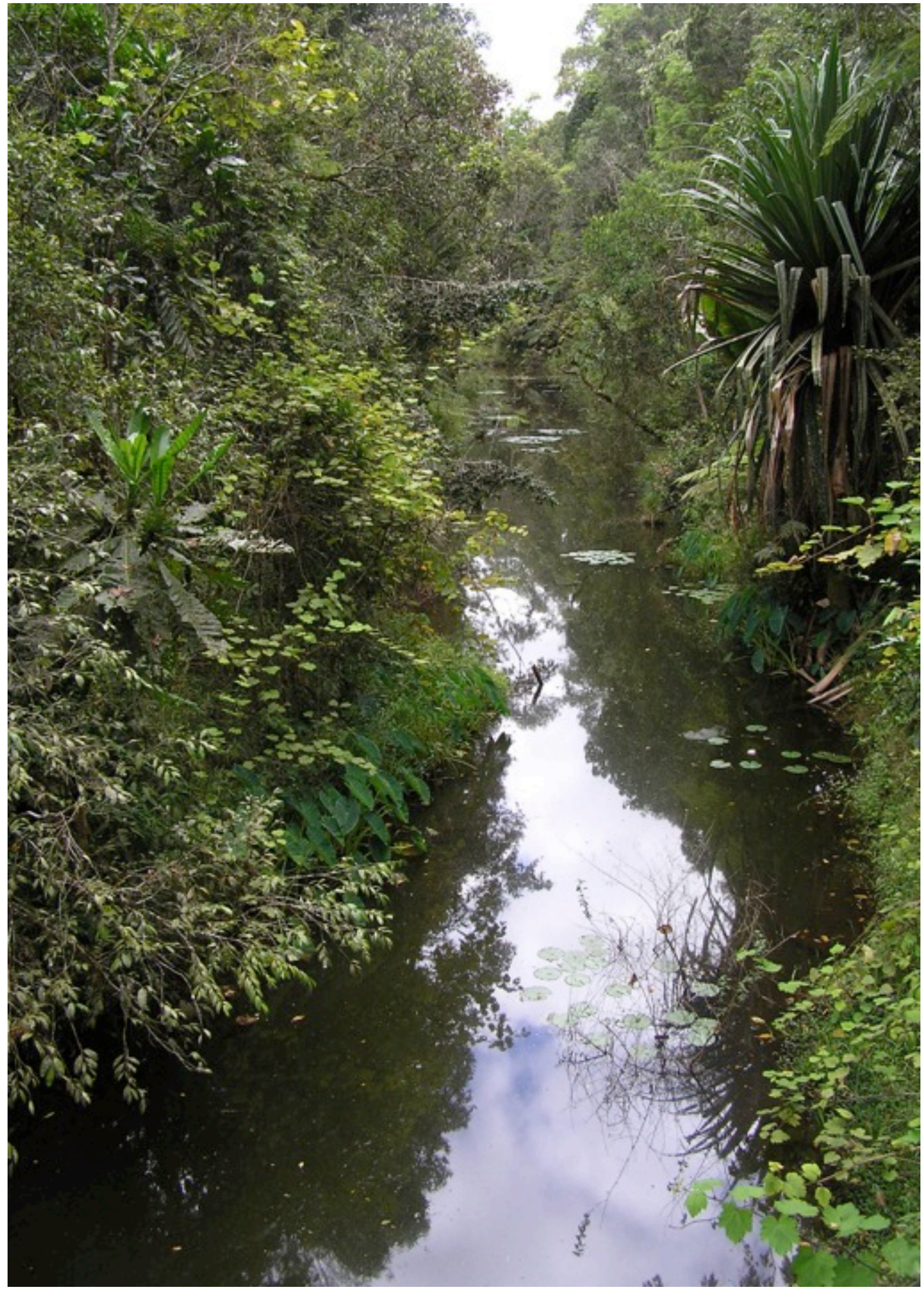
- **without adjustable parameters**
- **without pre-judging what specific mechanisms drive the system**

And thereby

- **gain insight into the forces that shape ecosystems**
- **make reliable predictions that can aid in conservation**

"It is interesting to contemplate
an entangled bank,
clothed with many plants of
many kinds,
with birds singing on the bushes,
with various insects flitting
about,
and with worms crawling
through the damp earth,
and to reflect that these
elaborately constructed forms,
so different in each other,
and dependent on each other in
so complex a manner,
have all been produced by
laws acting around us."

Charles Darwin
concluding paragraph of Origin
of the Species



Part 2: The Metrics of Macroecology:

- Meaning and uses of Metrics
- Mathematical representations
- Prevailing patterns

Macroecology:

The study of the distribution, abundance, energetics, and interaction network structure of individuals and species across multiple spatial and temporal scales

Census Data → **The Metrics of Macroecology:** *Where theory and data meet* ← **Ecological Theory**

↓

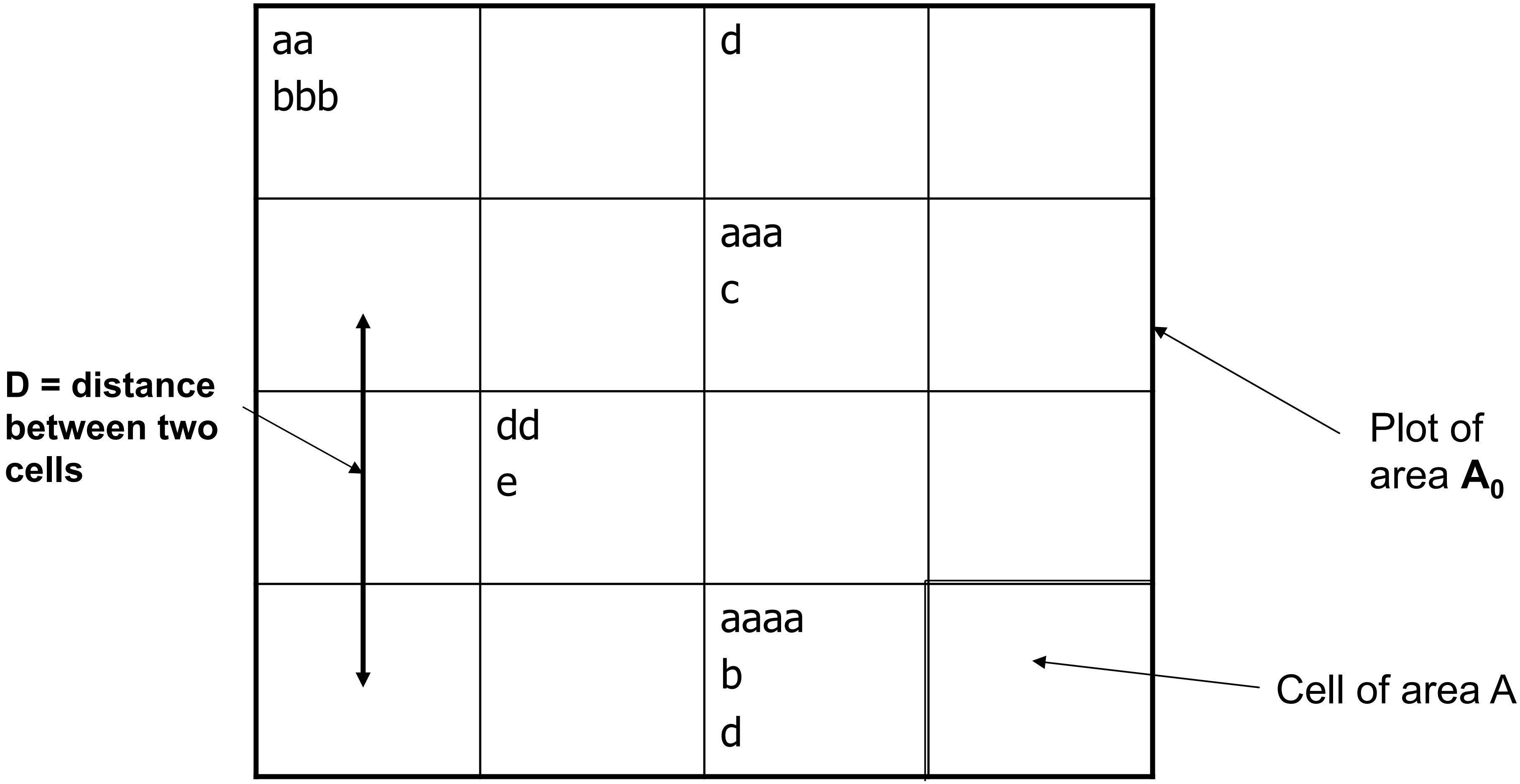
Applications to Global Change Science

- Improving climate envelopes for species
- Estimating species richness at large scales
- Loss of species and endemic species under habitat loss
- SAR and the problem of “over-stuffing the envelope”
- Network collapse under deletions
- Estimating rarity for conservation status
- Scaling up plot-scale results of manipulation experiments

Patterns & Metrics

1. # species increases with area censused.	Species-Area Relationship. (SAR)
2. Most species are rare, some abundant.	Species-Abundance Distribution (SAD)
3. Some individuals are big, most small.	Individuals Size Distribution
4. Common species have small individuals.	Size-abundance distribution
5. Individuals in species tend to aggregate.	Spatial-Abundance Distribution
6. More trophic specialists than generalists.	Foodweb node-linkage distribution

Scaling Metrics and Patterns

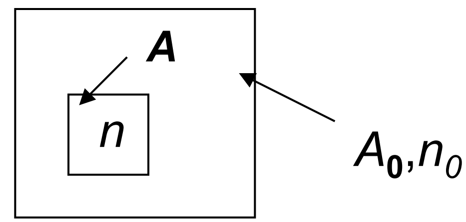


$\mathbf{S_0}$ (= 5) species, $\mathbf{N_0}$ (= 19) individuals

Metrics of Spatial Pattern in Ecology

<u>Species-Level</u> Properties: defined for a species with n_0 individuals in A_0

$\Pi(n|\ A, n_0, A_0)$ spatial abundance dist. — probability that n individuals are in a cell of area A



$C(A, D|n_0, A_0)$ occupancy correlation — probability the species is found in two cells of area A a distance D apart

(***with this notation we are anticipating that the only species trait that will matter here is n_0: we can test that***)

Metrics of Community Patterns in Ecology

<u>Some Community-Level</u> Metrics:

$S(A\|S_0,N_0, A_0)$	species-area reln. (SAR)	Expected number of species found in cell of area A
$E(A\|S_0,N_0,A_0)$	endemics-area reln. (EAR)	Expected number of species unique to cell of area A
$\Phi(n\|S_0,N_0)$	spec. abund. distribution (SAD)	Fraction of species with n individuals
$\Psi(\varepsilon\|S_0,N_0,E_0)$	energy distribution with	$\Psi d\varepsilon$ = Fraction of individuals in community with metabolic energy in $(\varepsilon, \varepsilon + d\varepsilon)$

NESTED SARs and EARs:

AAAA BB C D	A B C E
A D	AAA BBB

A_0 cell divided into 4 Cells

1. ***Species-Area Relationship***

   $S(A_0) = 5$; $S(A_0/4) = (4+4+2+2)/4 = 3$

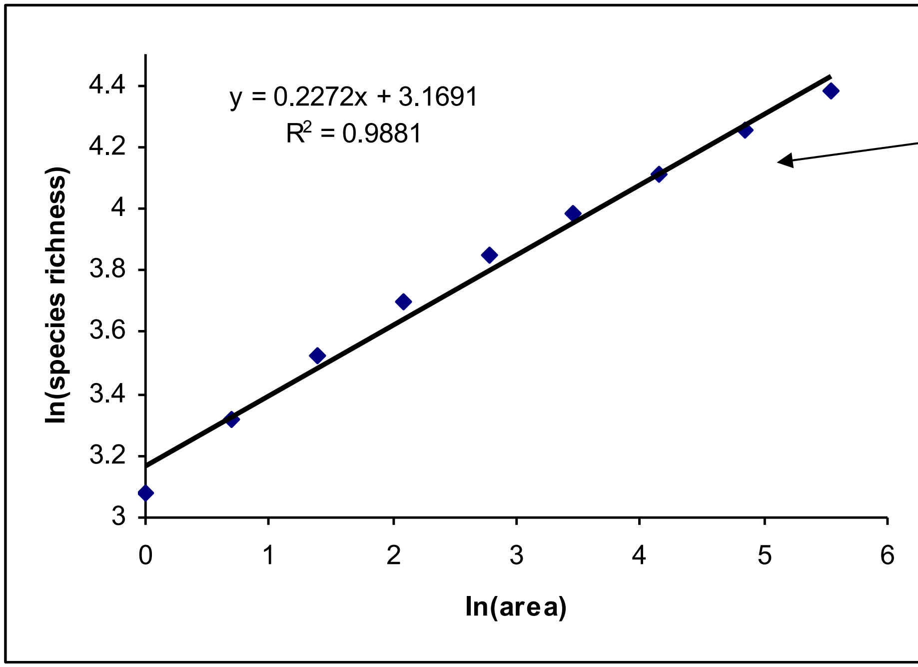
2. ***Endemics-Area Relationship***

   $E(A_0) = 5$; $E(A_0/4) = (0+1+ 0+ 0)/4 =1/4$

   sp. E

Pervasive Patterns Continued

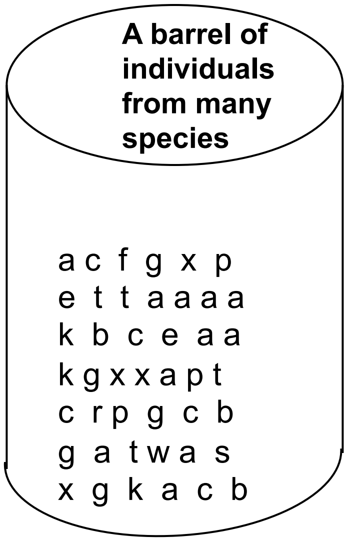
Over large scale ranges, power-law species-area relationships are the exception, not the rule.



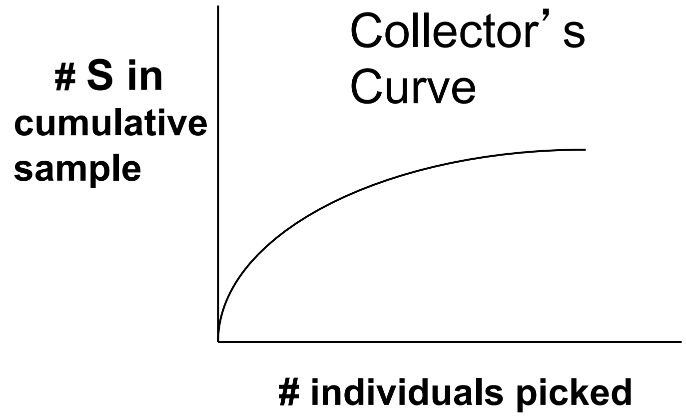
R^2 ~ 0.99 but clearly there is curvature!

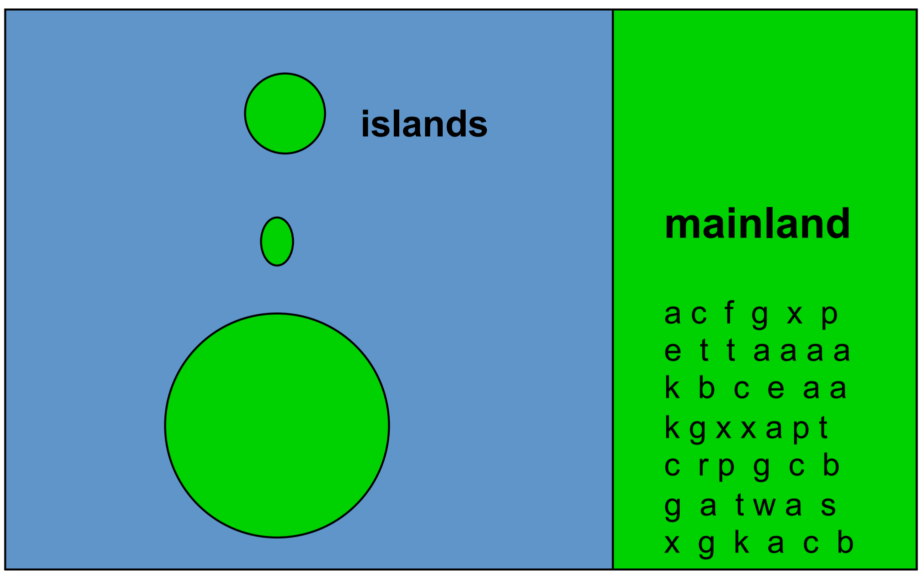
Scaling need not be fractal/power-law to be interesting, however.

A note on SARs and Collector's Curves



Grab individuals one at a time and note what species they are





Preston/May: . The canonical lognormal and the power-law SAR.

If the individuals on the islands are a random draw from the mainland, and the abundance distribution on the mainland is the canonical lognormal, then the collectors curve will be a power law with a slope of ~ 0.25.

If island abundance is proportional to island area, then the species area curve across islands is $S = c\,A^{0.25}$

These "island SARs" are really collector's curves, by the assumption that islands contain a random draw from the mainland. In contrast, our interest here is in mainland <u>nested</u> SARs

Additional Macroecological Metrics

- Distribution of number of trophic links per species in a food web
- Distribution of flow rates across the links in a food web
- Distribution of home range sizes
- Species range-size versus abundance relationship
- Distribution of metabolic energy rates or body sizes across
  - all the individuals in the community,
  - all the species, averaged over individuals in species,
  - all individuals within each species
- Distribution of dispersal distances

Can you think of others?

Three Related Scaling Patterns in Ecology:

The energy-equivalence rule: The total metabolic energy requirement of all the individuals in a size cohort is independent of the abundance of that cohort

The Damuth rule: abundance of species scales as $m^{-3/4}$

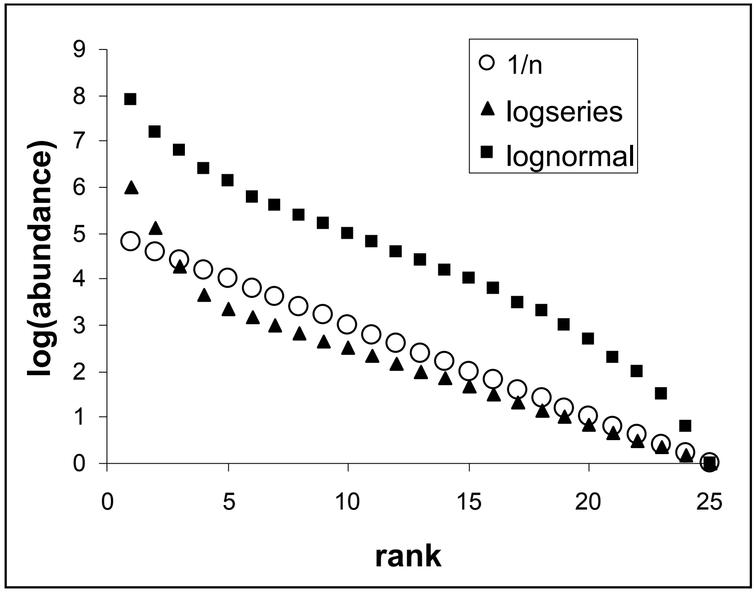
(m is average mass of individuals in species)

The metabolic scaling rule: metabolism of individuals scales as $\text{mass}^{3/4}$

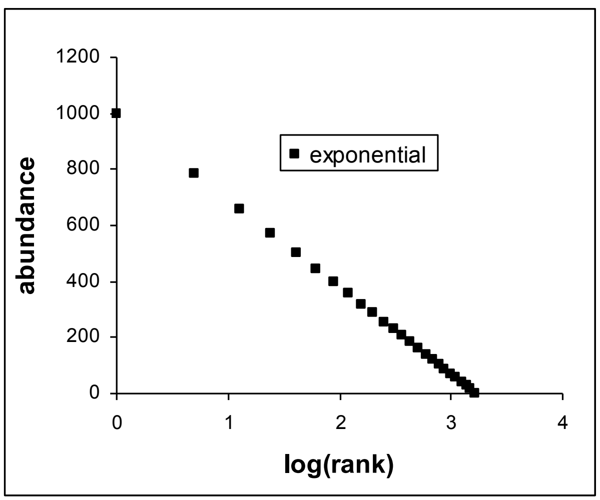
The Species-Abundance Distribution

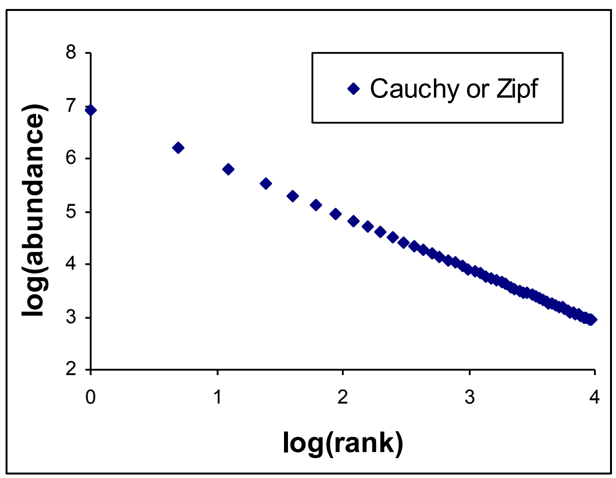
The distribution of abundances, $\{n_0\}$, across the species in an ecosystem is <u>generally consistent</u> with Fisher's log-series distribution $\Phi(n_0) \sim \exp(-kn_0)/n_0$, and often k is sufficiently small so that $\Phi(n_0) \sim 1/n_0$ (power law with exponent = -1)

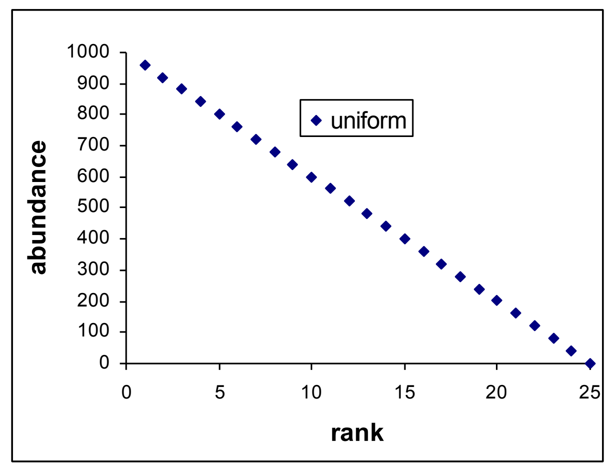
By plotting rank vs. abundance with different choices for axes, the pattern of the SAD can often be discerned.



(In some data sets, $\Phi(n)$ appears to be more lognormally distributed, however.)







Relationships among Metrics

$$S(A) = S_0 \sum_{n_0} [1 - \Pi(0|n_0, A, A_0)]\ \phi(n_0)$$

Probability a species with n_0 individuals in A_0 is present in A

Probability a species has n_0 individuals in A_0

$$E(A) = S_0 \sum_{n_0} \Pi(n_0|n_0, A, A_0)\ \phi(n_0)$$

Probability a species with n_0 individuals in A_0 has all its n_0 individuals in A

Useful mathematical relationships

1. Converting probability densities to new independent variables. ***We are given a probability density function f(x) and another variable, y, which can be expressed as a function of x: y = y(x). Inverting y = y(x), we can also write x = x(y). For example, if y(x) = x^2, then x(y) = √y. The probability distribution for y, g(y), is given by***

$$g(y) = f(x(y))\,dx/dy$$

The term d*x*/d*y* is needed to ensure that if we integrate each distribution over equivalent ranges of their independent variable we get the same result: ∫d*x* f(*x*) = ∫d*y* *g*(*y*).

2. Deriving a Dependency Relation from a conditional probability distribution

$$\bar{\mathrm{z}}(\mathrm{x}) = \sum_{\mathrm{z}} \mathrm{z} \cdot \mathrm{p}(\mathrm{z} \mid \mathrm{x}).$$

3. Going back and forth between a probability distribution, *Φ*(*n*), and a rank-abundance relationship, *n*(*r*).

$$\Phi(n) = \frac{-1/S_0}{dn/dr}$$

Why do we care about patterns and metrics in ecology?

1. **Extinction rates under habitat loss.**

   25% of Amazon rain forest has been cut.
   How many species lost?

2. **Scaling up biodiversity.**

   How many species of arboreal beetles
   in all of the Amazonian rain forest?

3. **Inferring process from pattern**
   (analogy: *Brahe -> Kepler -> Newton*)

Data $\longrightarrow$ **Pattern** $\longrightarrow$ **Force Law**

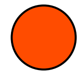
3. The MaxEnt Method

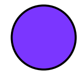
- Thermodynamic and Information Entropies
- Outcomes of MaxEnt
- Past applications
- What if it doesn’t work?
- MaxEnt and the logic of inference
- History of the concept: the Laplace urn problem.

A Little History: The Laplace Problem

An urn with red and blue balls in it.

Initially in urn:

 # = **?**

 # = **?**

Suppose initially, you have pulled out (with replacement)

R red balls and B blue balls.

The probability the next one will be red is:

P(R+1|R,B) = (R+1)/(R+B+2)

This is called the Laplace Rule of Succession

Maximum Entropy?
Just what is being maximized?

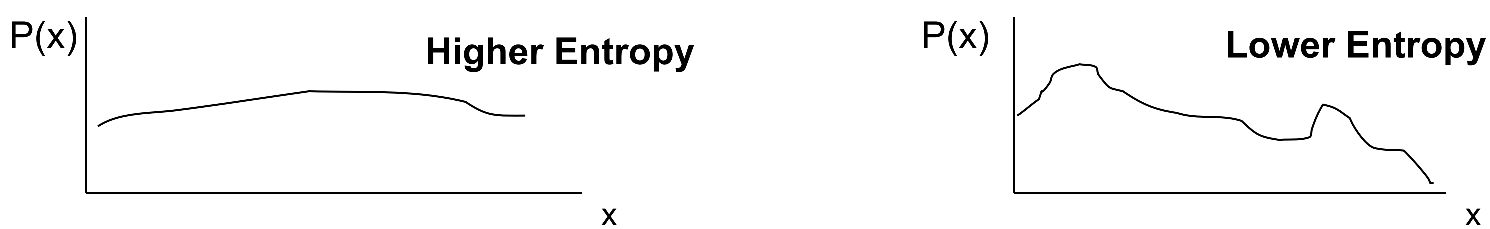
Here “**entropy**” refers to

information entropy,

not

thermodynamic entropy.

Information entropy is a measure of the lack of structure or detail in the probability distribution describing your knowledge of a system.



The Maximum Entropy (MaxEnt) Inference Procedure

1940's – 1960's: Claude Shannon, Edwin Jaynes

(The basic ideas go back to Laplace)

Suppose you wish to know the form of a probability distribution $P(n)$.

All you know is some prior information about $P(n)$ in the form of <u>constraints</u>.

For example: you might know the mean value, or the variance, or some combination of moments of the distribution.

What can you infer is the least biased form of the probability distribution?

The Maximum Entropy (MaxEnt) Inference Procedure

You seek the form of a probability distribution *P*(*n*).

All you know is some prior information about *P*(*n*): i.e., constraints.

What is your least-biased inference of the shape of *P*(*n*)?

Fundamental proven theorem: The *P*(*n*) that maximizes

$I = -\sum_n P(n) \log(P(n))$ **(*I* is "Information Entropy")**

with *P*(*n*) subject to prior constraints, is the least biased estimate of the shape of *P*(*n*).

"Least biased": smoothest possible distribution that satisfies the constraints Any other *P*(*n*) would implicitly incorporate additional information that you do not possess.

Jaynes, E. T., 1982, On the Rationale of Maximum Entropy Methods, *Proc. IEEE.*, 70, 939;

A "Derivation" of the main result

Constraints: $\sum_n f_k(n)p(n) = \langle f_k \rangle$ $\sum_n p(n) = 1$ (k = 1,…. K = # constraints)

Maximize: $S_I = -\sum_n p(n)\log(p(n))$ over the function $p(n)$

Approach: let $W = -\sum_n p(n)\log(p(n)) - \lambda_0(\sum_n p(n) - 1) - \sum_k \lambda_k(\sum_n f_k(n)p(n) - \langle f_k \rangle)$

(λ 's are Lagrange Multipliers)

$$dW/dp = -\log(p(n)) - 1 - \lambda_0 - \sum_k \lambda_k f_k(n) = 0$$

Solution:

$$p(n) = \frac{e^{-\sum_{k=1}^{K}\lambda_k f_k(n)}}{Z(\lambda_1, \lambda_2, \ldots, \lambda_K)}$$

Where:

$$Z = \sum_n e^{-\sum_{k=1}^{K}\lambda_k f_k(n)} \qquad \frac{\partial \log(Z)}{\partial \lambda_k} = -\langle f_k \rangle$$

Some **examples** of outcomes of the MaxEnt procedure:

Let the constraints be of the form: $F_k = \langle f_k(n) \rangle = \sum_n P(n) f_k(n)$

Constraint function $f(n)$	**Form** of $P(n)$
n	$e^{-\lambda n}$
n, n^2	Gaussian (normal) distribution
$\log(n)$, $\log^2(n)$	Lognormal distribution
$\log(n)$	$n^{-\lambda}$ (i.e., power law)
Discrete constraints, $P(n_i) = a$, can be handled as well	

Some past applications of MaxEnt:

1. **Improving image resolution in medicine, forensics (Skilling, …)**

2. **Inferring least-biased numerical values for gaps in economic data such as in input-output tables (George Judge, Amos Golan)**

3. **Deriving the laws of stat. mech./thermodynamics (Jaynes)**

4. **Improving estimation of climate envelopes for species (Elith, Phillips)**

What if MaxEnt gives a poor prediction?

1. **You made a mathematical error in working out the solutions to the equations.**

2. **Your prior knowledge (in the form of constraints) is not factually correct. If one of your constraints is the mean value over the sought distribution, and you mis-measured it or the values of your state variables were mis-estimated, your predictions will suffer as a consequence.**

3. **Your constraints may not provide enough adequate information to determine a good answer. If you neglect information, MaxEnt will do its best for you. But it may not be good enough. YOUR ANSWER WILL BE THE BEST POSSIBLE, GIVEN THE CONSTRAINTS THAT YOU USE.**

A General Rationale

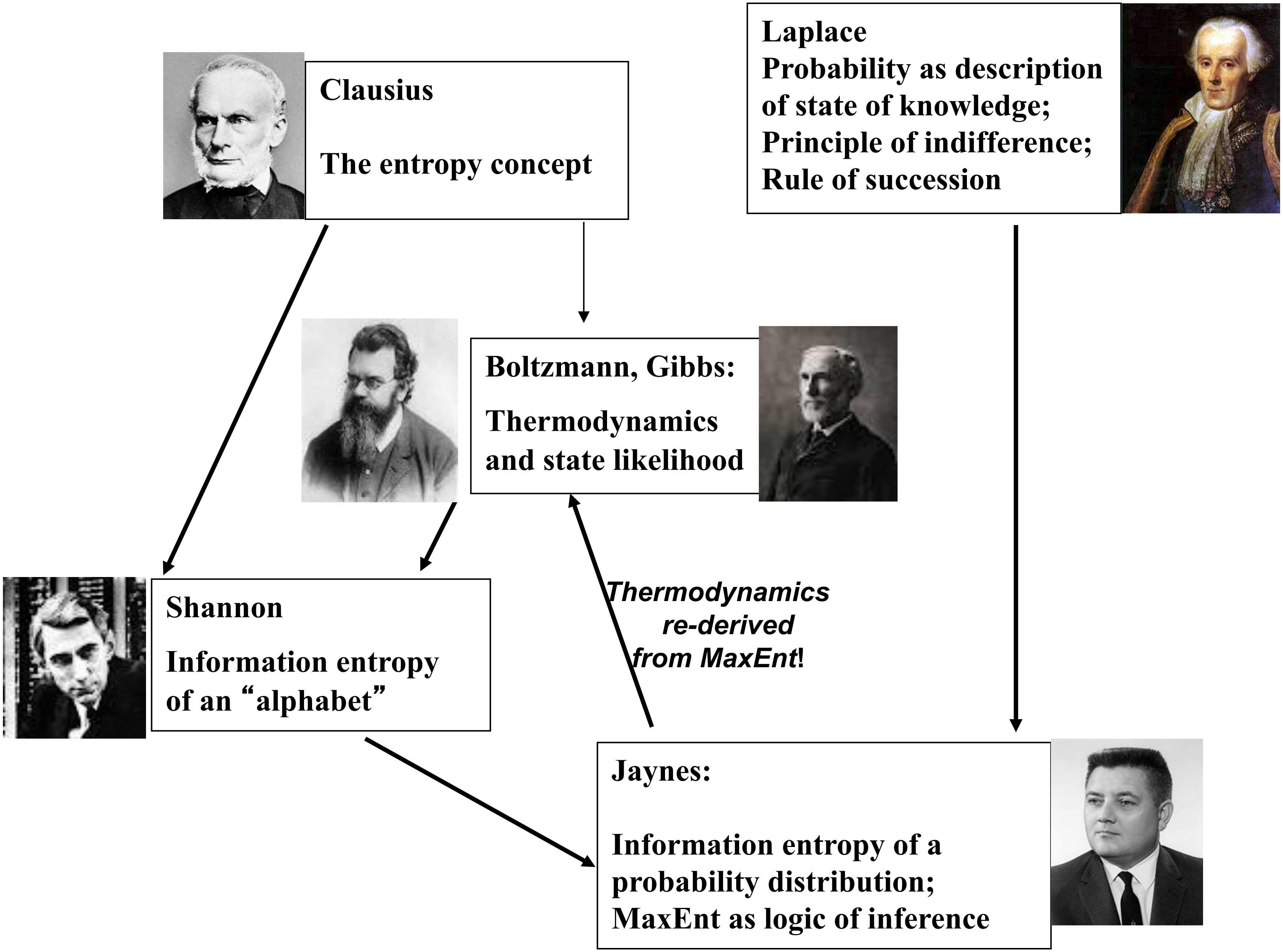
for the Use of the concept of information entropy in science:

1. In science we generally **begin with prior knowledge** and seek to expand that knowledge.

2. **Knowledge is not absolute, but rather probabilistic** in nature, and thus the expanded knowledge we seek can often be expressed mathematically in the form of presently unknown probability distributions.

3. Our **prior knowledge** can often be expressed **in the form of constraints** on those unknown distributions.

4. We seek **expanded knowledge that is "least biased",** in the sense that the expanded knowledge does not assume anything about the distributions other than the information contained in our prior knowledge.



Clausius
The entropy concept
Laplace
Probability as description of state of knowledge;
Principle of indifference;
Rule of succession
Boltzmann, Gibbs:
Thermodynamics and state likelihood
Shannon
Information entropy of an “alphabet”
Thermodynamics re-derived from MaxEnt!
Jaynes:
Information entropy of a probability distribution;
MaxEnt as logic of inference

4: The Maximum Entropy Theory of Ecology

- The State variable Concept
- Definition of the Core Distributions
- Predictions of the Theory

The Goal: Predicting the Form of the Metrics of Macroecology

- ***S*pecies-area relationship (SAR)**

- **Species-abundance distribution (SAD)**

- **Spatial distribution of individuals**

- **Linkage distribution across Nodes in Food Web**

- **Metabolic rate and body-size distributions over individuals and species**

- **… and many others**

MaxEnt and the State Variable Concept

In **Thermodynamics**, these state variables characterize the system:

P: pressure

V: volume

T: temperature

n: number of moles

PV=nRT, Boltzmann distribution of energy levels, entropy law, equipartition, binomial distribution of molecules in space … can all be derived from MaxEnt, with constraints provided by these state variables. (Jaynes 1957a, b)

In **Ecology** we start with:

Harte et al. (2008) *Ecology* 89:2700-2711;

(2009) *Ecology Letters* 12: 789-797

Harte: Oxford U. Press: June 2011

"Maximum Entropy and Ecology"

A_0 *: area of ecosystem or census plot*

S_0 *: total number of species in* A_0

N_0 *: total number of individuals amongst all those species*

E_0 *: total metabolic rate of all those individuals*

And show that from the constraints their ratios impose we **can use MaxEnt to predict the metrics of macroecology**

(**without any adjustable parameters…no data fitting**)

The Maximum Entropy Theory of Ecology (METE)

Two probability distributions comprise the theory:

1. $$R(n, \varepsilon | A_0, S_0, N_0, E_0)$$

abundance

metabolic
energy rate

State
variables

R is defined over the species and individuals in an area A_0.

$R \cdot d\varepsilon$ = probability that if a species is picked from the species pool, then it has abundance n, and if an individual is picked at random from that species then its metabolic energy requirement is in the interval $(\varepsilon, \varepsilon + d\varepsilon)$

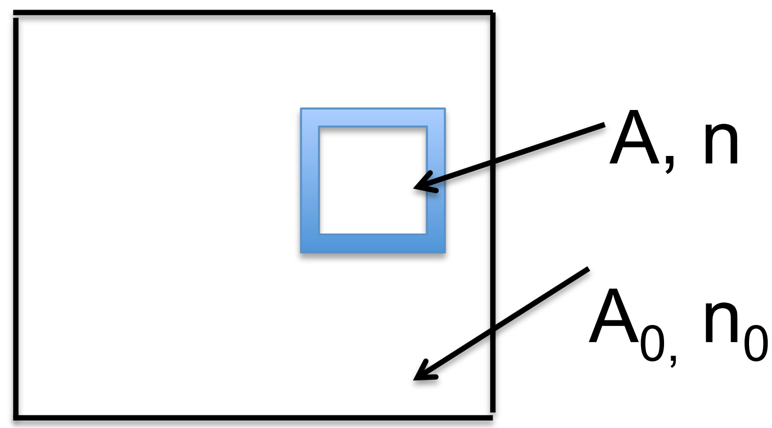
.

2. … and a species-level spatial distribution,

$$\Pi(n|A,n_0,A_0)$$

describing aggregation of individuals within species:

Π =
probability
that n individuals
in A if n_0 in A_0



From R and Π, most of the metrics
of macroecology can be derived.

METE gives

$$R(n,\varepsilon \mid S_0, N_0, E_0) = \frac{e^{-\lambda_1 n} e^{-\lambda_2 n \varepsilon}}{Z(\lambda_1, \lambda_2)}$$

(Summing over n)

$$\Psi(\varepsilon \mid S_0, N_0, E_0) \approx \lambda_2 \cdot \beta \cdot \frac{e^{-\gamma}}{(1-e^{-\gamma})^2}$$

Distribution of metabolic rates over individuals

$$\Theta(\varepsilon \mid n) = \frac{R}{\Phi} = \lambda_2 n e^{-\lambda_2 n(\varepsilon - 1)}$$

Intraspecific metabolic rate distribution

(Taking the mean)

$$< \varepsilon(n) >= 1 + \frac{1}{n\lambda_2}$$

Abundance-metabolism relation for species

(Integrating over ε)

$$\Phi(n) = \frac{1}{\lambda_2 Z n} e^{-\beta n}$$

Species Abundance distribution

$$\Pi(n \mid n_0, A, A_0) = \frac{e^{-\lambda_\Pi n}}{Z_\Pi}$$

Species-level Spatial distribution: probability *n* individuals in a cell of area *A* if n_0 in A_0

$$S(A) = S_0 \sum_{n_0=1}^{N_0} \Phi(n_0) * [1 - \Pi(0 \mid n_0, A, A_0)]$$

The Species-Area relationship

$$E(A) = S_0 \sum_{n_0=1}^{N_0} \Phi(n_0) * \Pi\ (n_0 \mid n = n_0, A, A_0)$$

The Endemics-Area relationship

No derivations here: you have to trust me on the math…or do it yourself.

There are no adjustable parameters: the state variables uniquely determine the metrics

5. Tests of Predictions

At ~ 20 distinct habitats: ~ 2×10^4 Species, 5×10^7 individuals

36 serpentine meadow plots in CA

11 Smithsonian tropical forest plots

A 9.8 ha dry forest plot at San Emilio, Costa Rica

Plant census in Anza Borrego desert

Breeding bird censuses in southern Africa

Forest floor vegetation from 15 m^2 plots subalpine temperate montane forests

Tree census data from the Western Ghats in S. India

Early successional data from a massive earthflow event in the Rockies

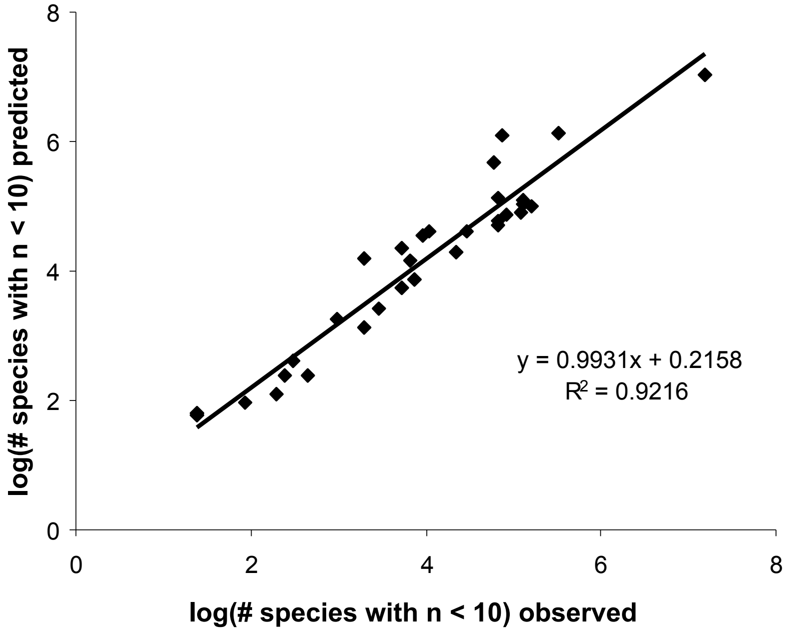
Hawaiian arthropods

Panamanian arthropods

...

Examples of validated predictions

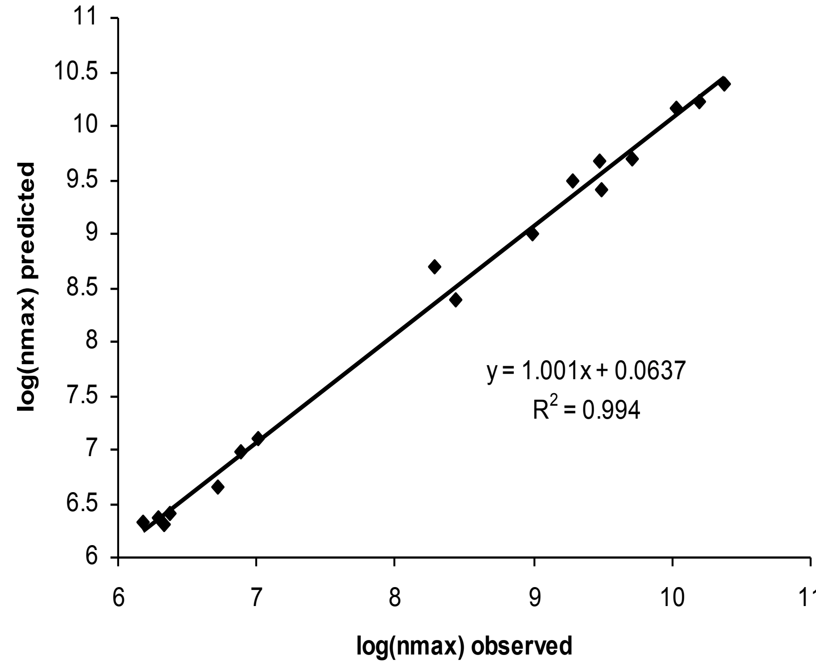
MaxEnt predicts the number of rare species



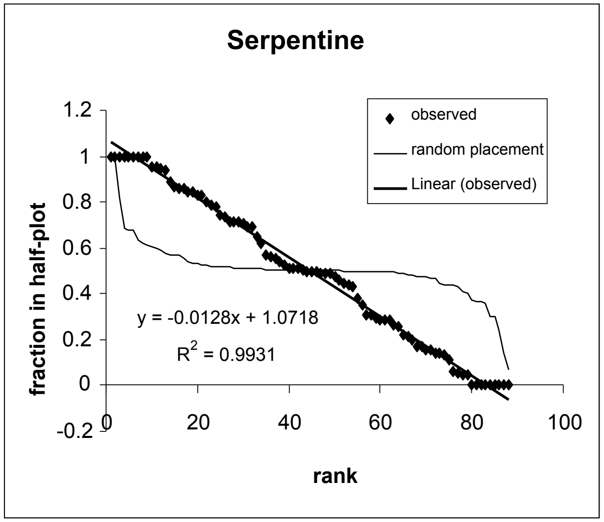
Harte et al., Ecology, 2009

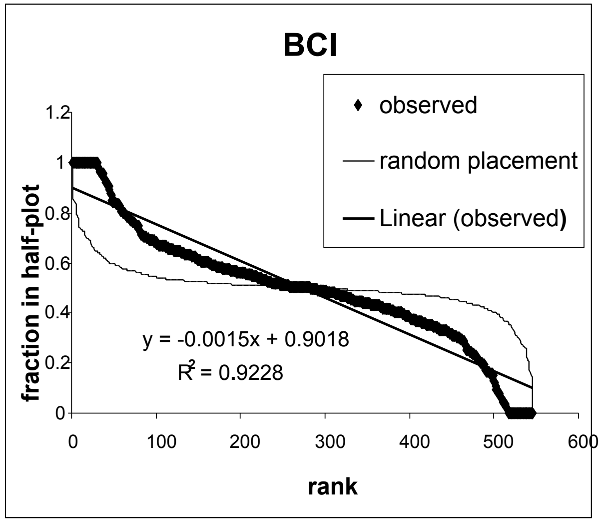
Harte, Oxford U. Press, 2011

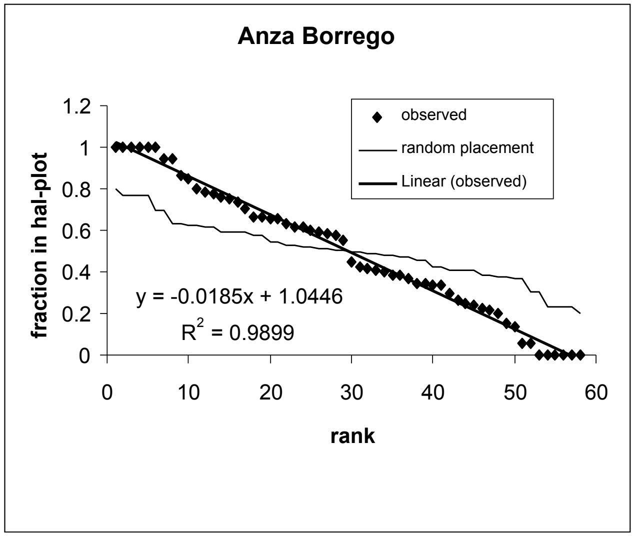
… and the abundance of the most abundant species

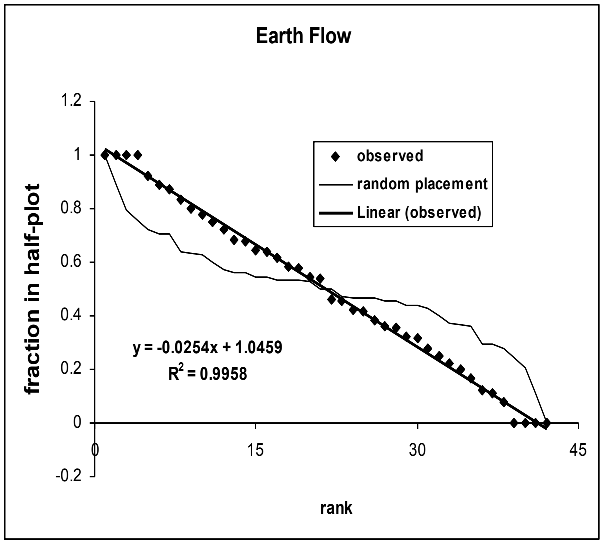


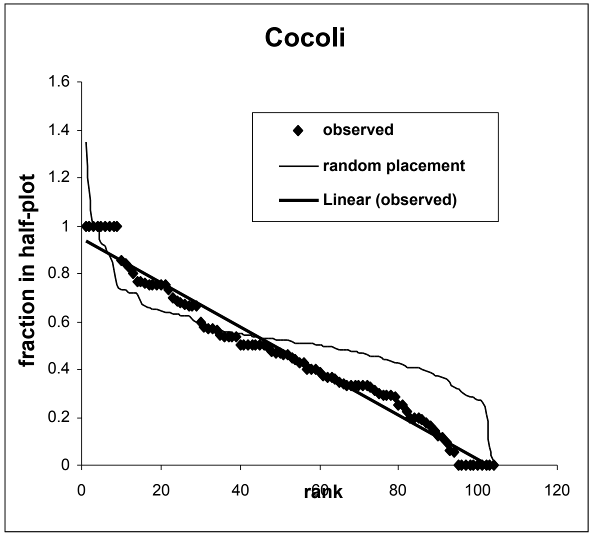
Bisection Graphs





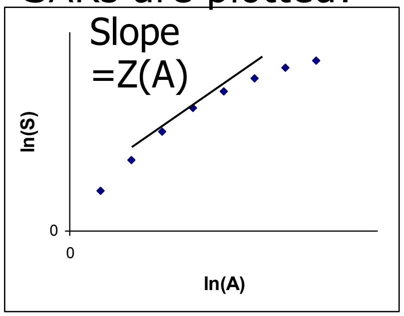




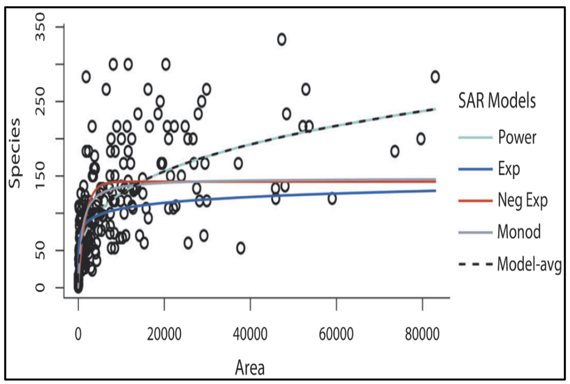


An Example of a Validated Prediction: The SAR

The traditional way SARs are plotted:



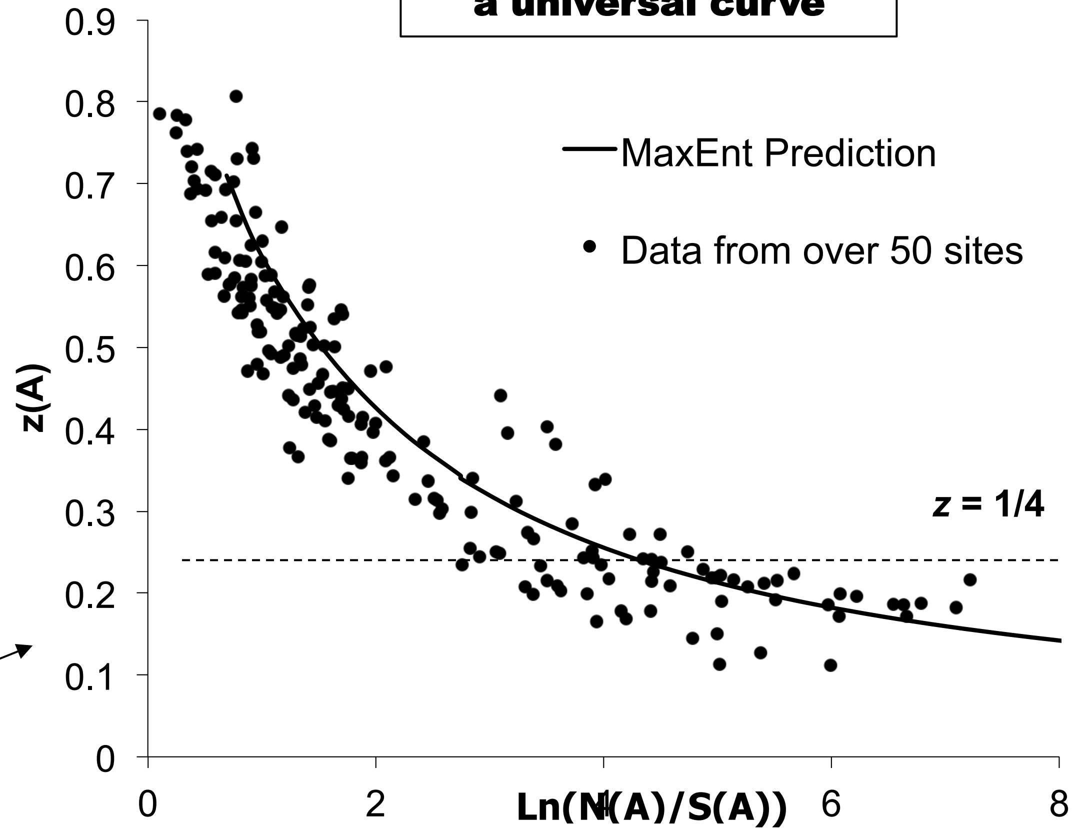
The mess that results:



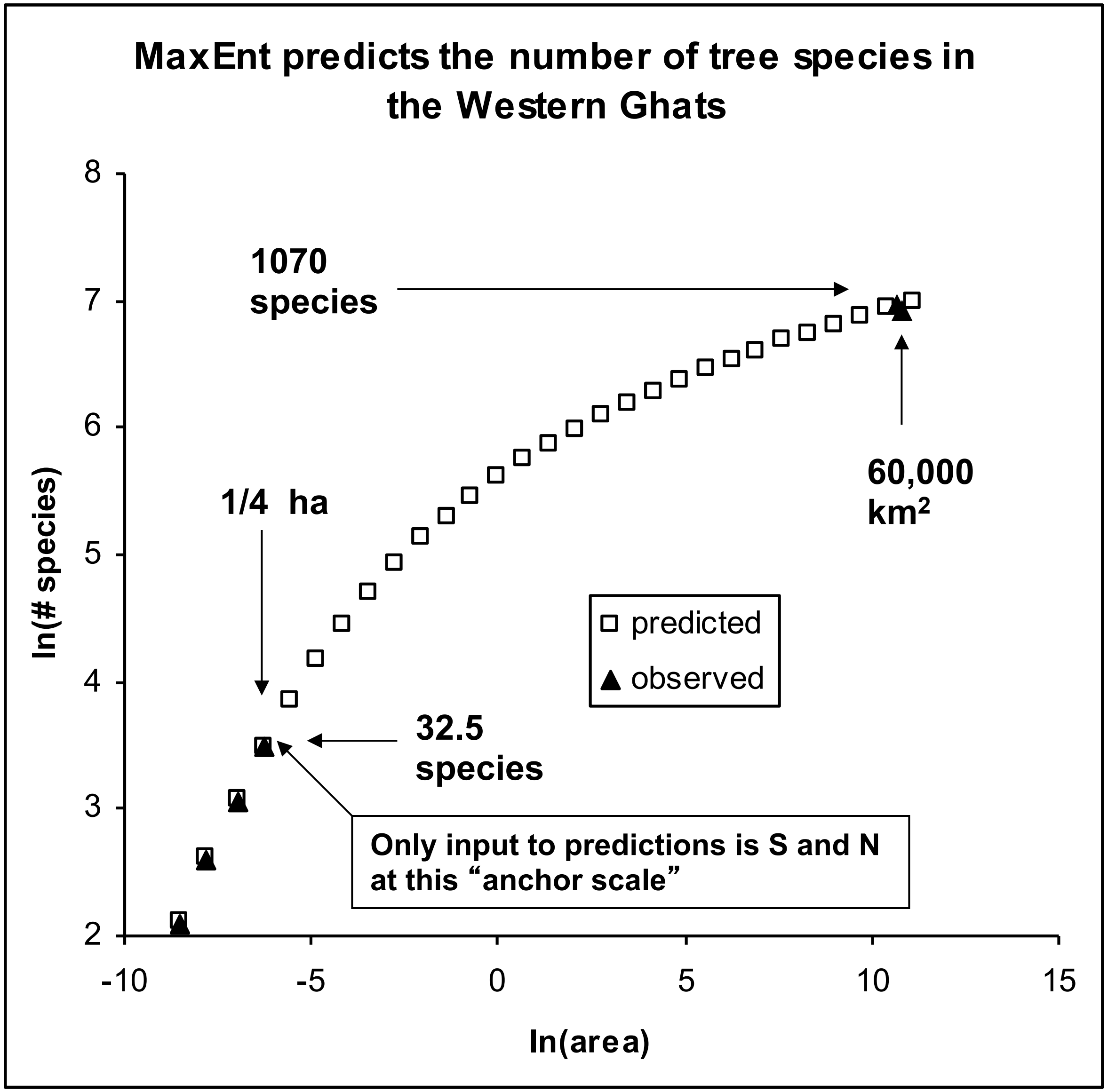
Re-plot to exhibit universality

MaxEnt predicts:

all species-area curves collapse onto a universal curve



Harte et al., Ecology Letters, 2010;
Harte, Oxford U. Press, 2011



Downscale from "anchor value"

Upscale from "anchor value"

We are extrapolating here over a scale range of

60,000 km^2 / ¼ ha

= 2^{24} !!

Presently, approximately 990 tree species are listed for the region, but that is certainly a low estimate of the true number.

Up-scaling Tropical Diversity Data

Taxa	**Plots (0.04 ha) Measured**	**SLPA (6000 ha) Predicted**	**All Panama (~8 x10^6 ha) Predicted**	**All Panama** Power law prediction ($z = ¼$)	**Amazonia (6 x 10^8 ha) Naïve prediction**	**Amazonia (by subregion) Realistic prediction**
Arthropods	1530	22,500	43,550	182,000	80,000	320,000
Trees	47	1180	2280	5600	3000	16,000
NOTES		2 x Basset et al.; ½ of Terry Erwin estimate	Condit (2010) estimates **2300** obs. tree species;.	Way too high	Way too low	subregions of non-overlapping species

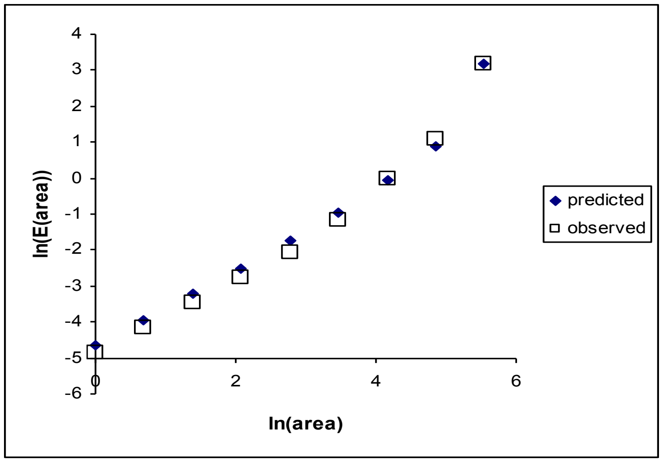
Let M = # distinct subregions in Amazonia, and assume for simplicity they are of comparable area and species richness:

Then $S(A_{\text{Amazonia}}) = M*S(A_{\text{Amazonia}}/M)$

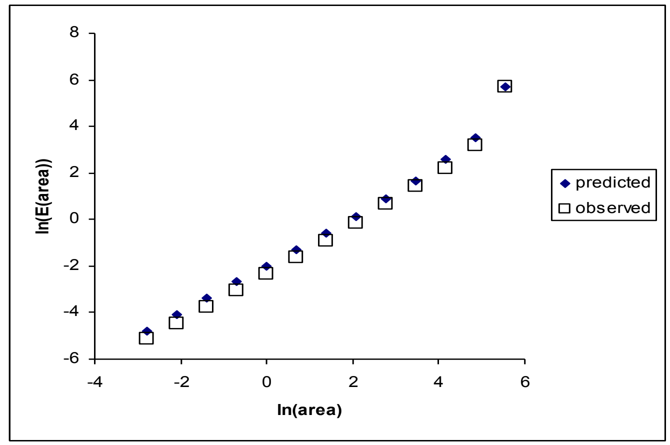
16,000 trees species are known and this results in M ~ 6. That in turn results in ~320,000 arthropod species.

The Endemics-Area Relationship

$$E(A) = S_0 \sum_{n_0=1}^{N_0} \Phi(n_0) * \Pi \ (n_0 \mid n = n_0, A, A_0)$$



Serpentine



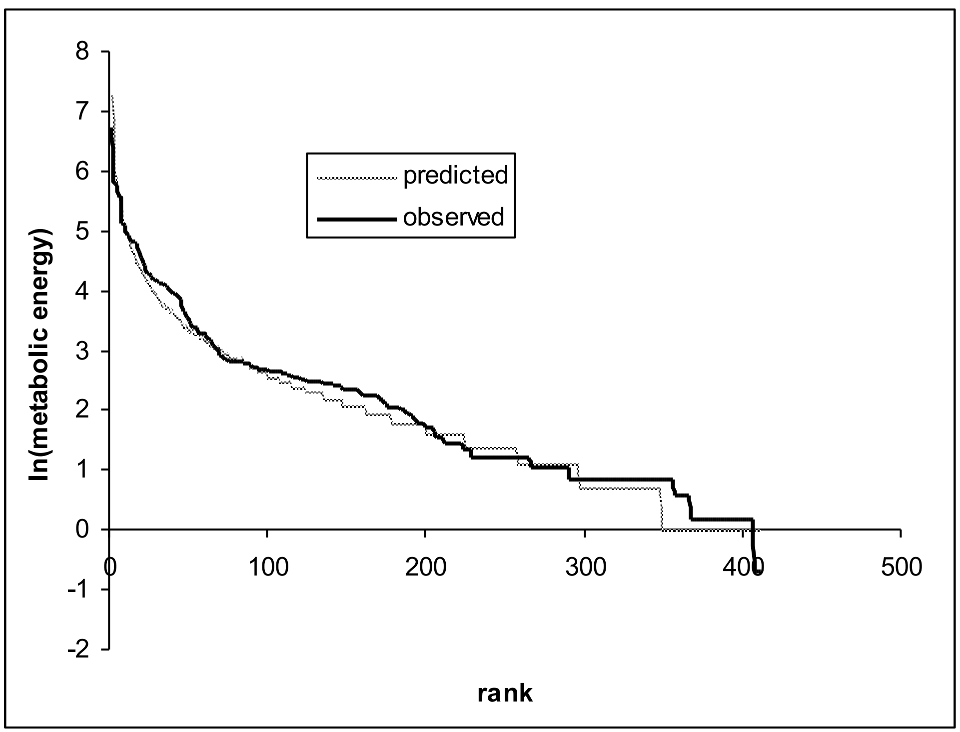
BCI

Tests of the energy distribution

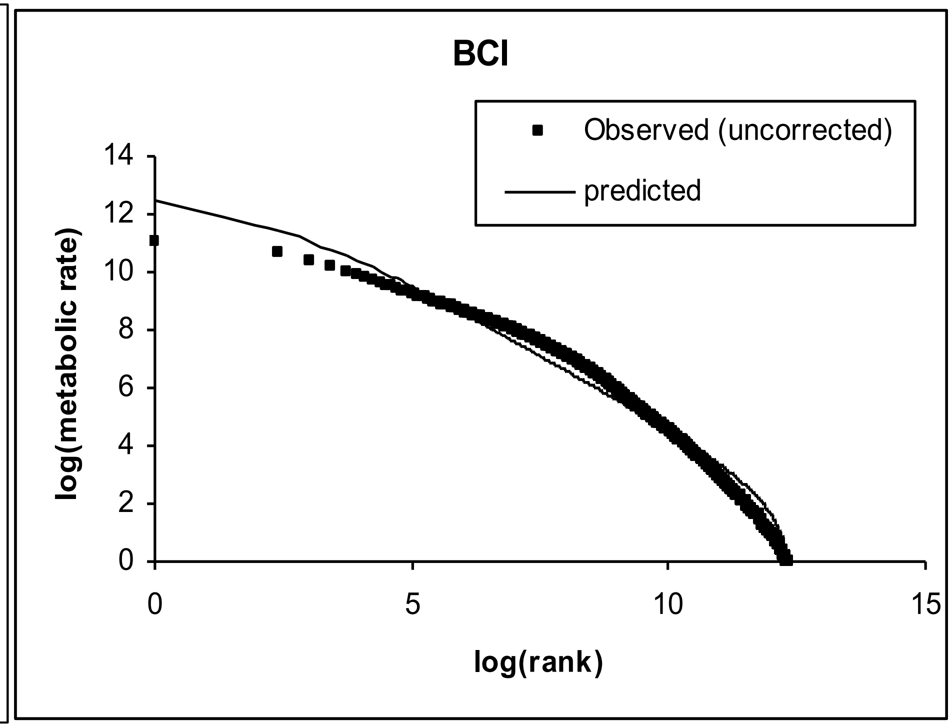
$$\Psi(\varepsilon) = \frac{\lambda_2 \beta\, e^{-\gamma}}{(1 - e^{-\gamma})^2}$$

$\gamma = \lambda_1 + \lambda_2 \varepsilon$

pit-trapped insects; assume $\varepsilon \sim m^{3/4}$



BCI forest; assume $\varepsilon \sim$ basal area



6. At the Frontier of METE

Core theory

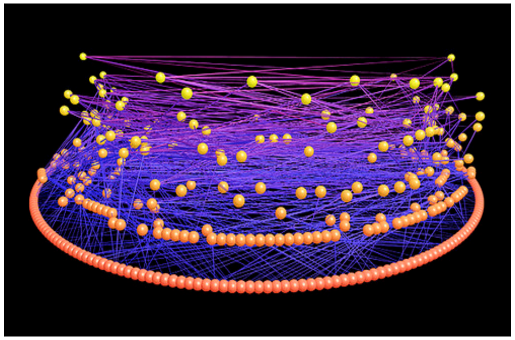
S, N, E

Resource constraints:

Water, **P**hosphorus,..

Trophic interaction constraints:

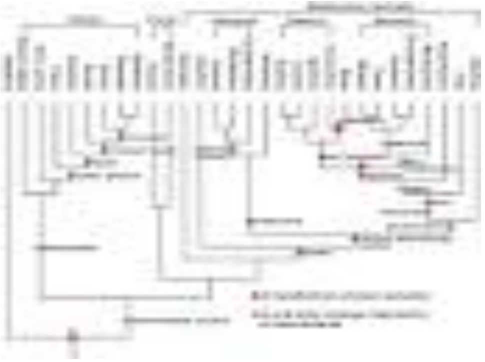
Linkages



Evolutionary constraints:

taxonomy/ phylogeny

Order, **F**amily, **G**enus

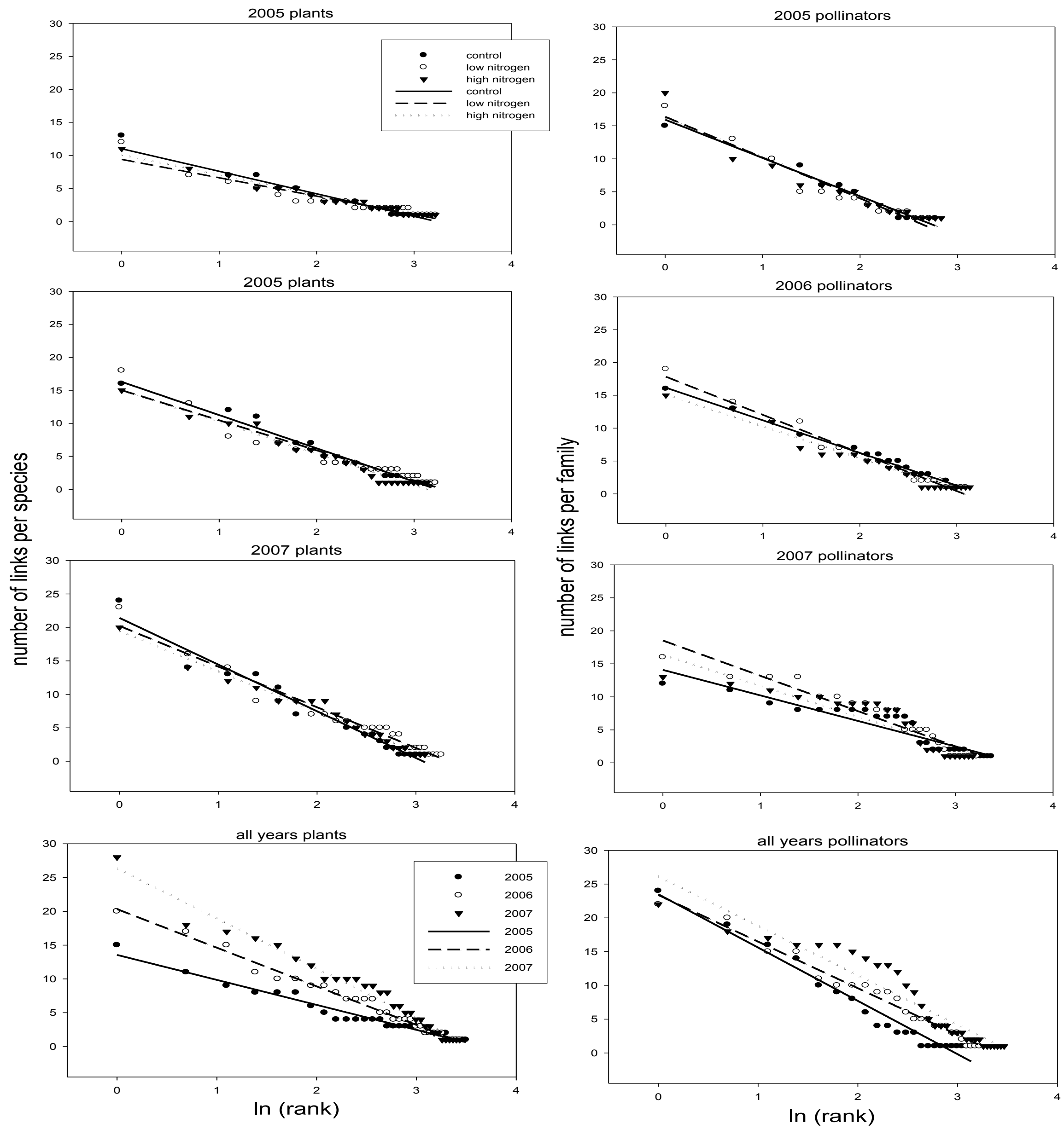


MaxEnt and Food Web Structure

Add state variable L (total number of linkages in web),

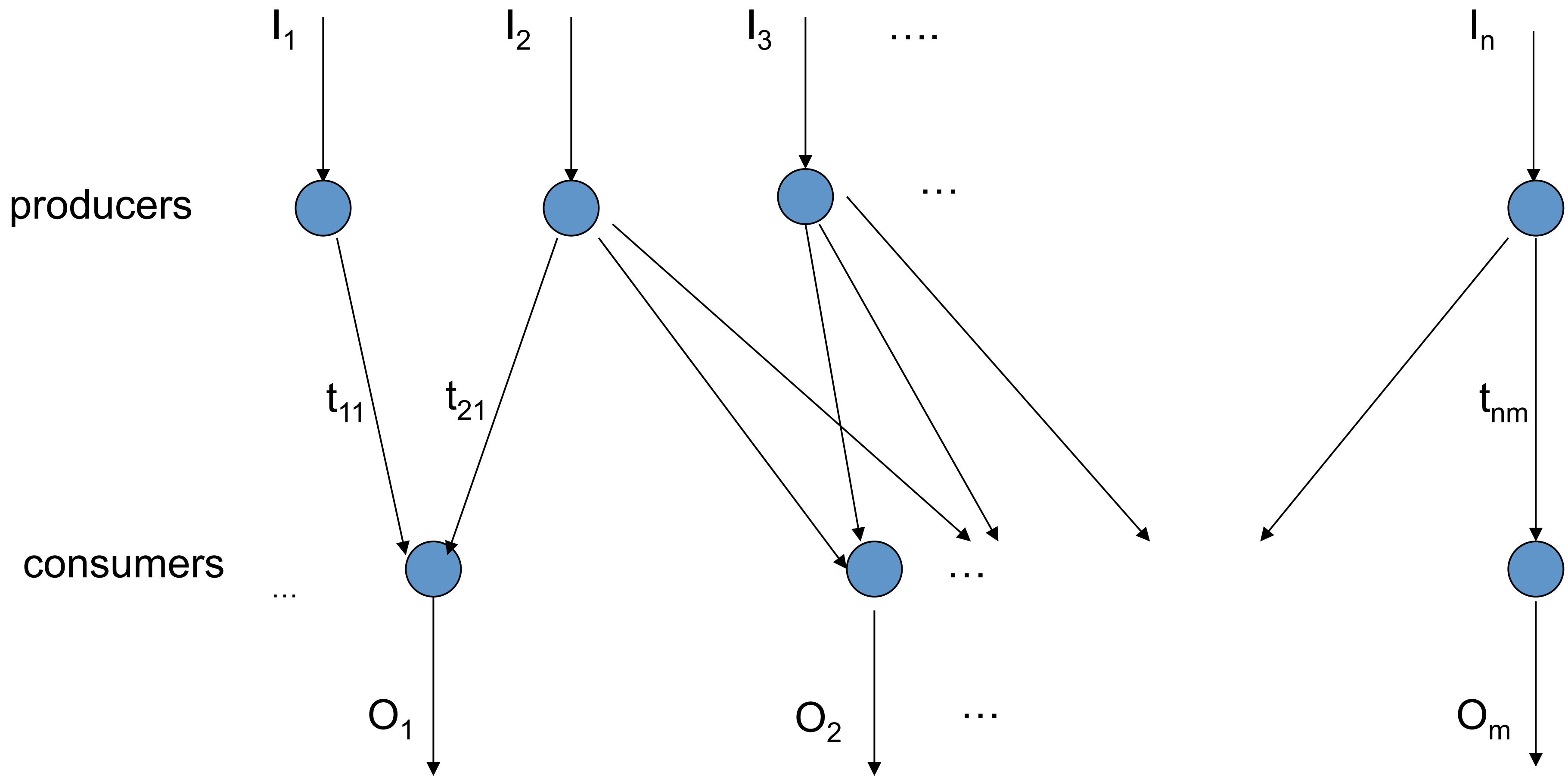
And MaxEnt predicts the distribution of:

 # linkages per species



See also Williams, 2010

FLOWS IN NETWORKS: The MaxEnt Solution

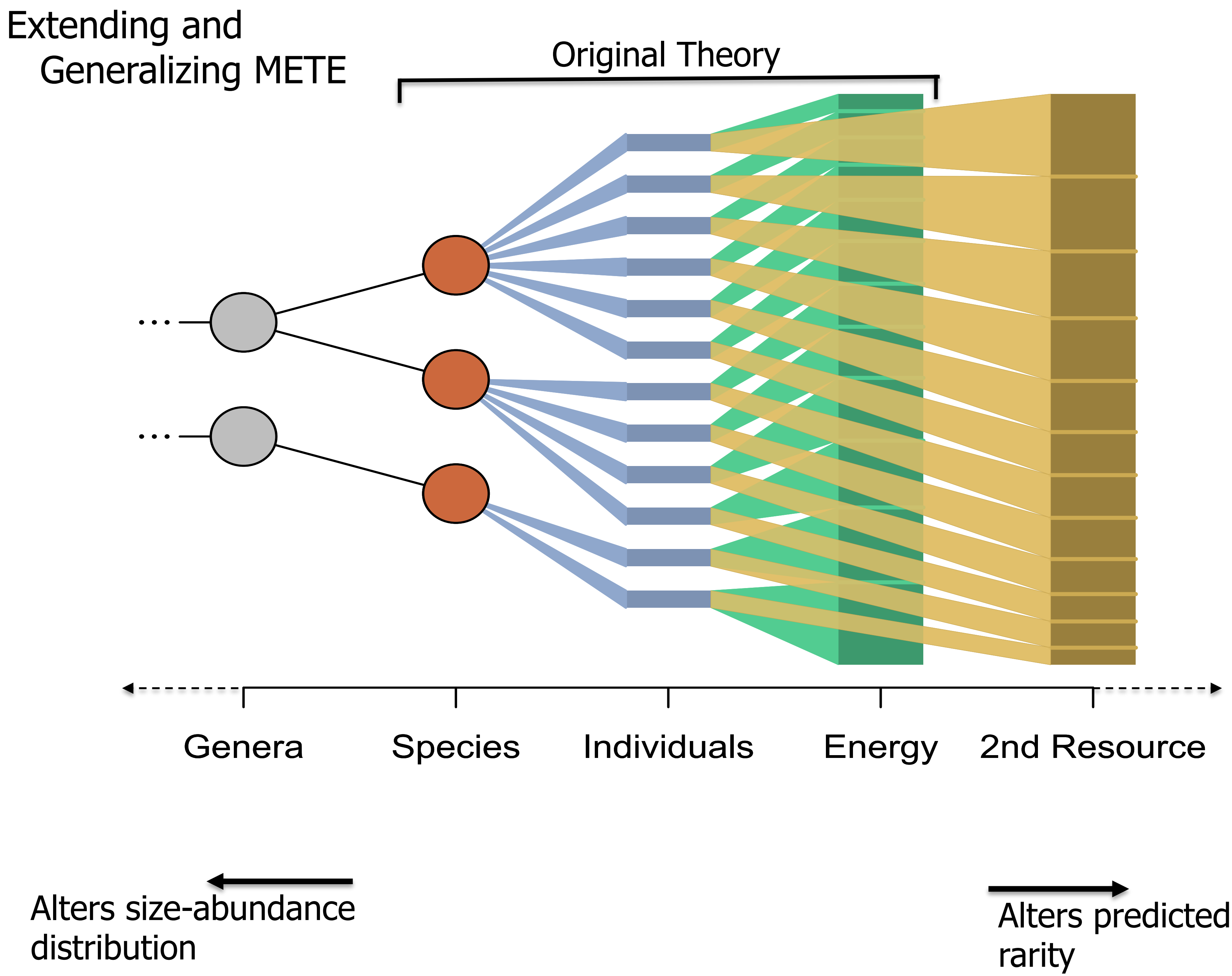


$T = \Sigma_{ij} t_{ij}$; $\mathbf{t_{ij} = I_i * O_j / T}$

E. G.:

The I's are NPP's of plants

The O's are metabolic rates of herbivores



Extending and
Generalizing METE
Original Theory
Genera
Species
Individuals
Energy
2nd Resource
Alters size-abundance
distribution
Alters predicted
rarity

Including additional resource constraints (in addition to energy, *E*)

The log-series SAD becomes:

$$\Phi(n) \sim \frac{e^{-\lambda n}}{n^{r}}$$

***r* - 1** = # additional resources

The inclusion of additional resource constraints predicts increased rarity

Extension of METE to higher taxonomic levels

Example: inclusion of genus as a category

State Variables:

G_0 = # genera
S_0 = # species
N_0 = # individuals
E_0 = total metabolic rate

The probability function *Q* replaces *R*

$Q(m,n,\varepsilon|G_0, S_0, N_0, E_0)$, defined as follows:

Pick a genus;
***Q* is the probability it has *m* species**
and if you pick one of those species from that genus,
that it has *n* individuals,
and that if you pick one of those individuals from that species,
that it has metabolic rate ε.

The constraints:

$$< m > = \frac{S_0}{G_0} = \sum_{m,n,\varepsilon} mQ$$

$$< n_G > = \frac{N_0}{G_0} = \sum_{m,n,\varepsilon} mnQ$$

$$< \varepsilon_G > = \frac{E_0}{G_0} = \sum_{m,n,\varepsilon} mn\varepsilon Q$$

Now we can predict the “old” metrics that the S,N,E theory predicts:

Species abundance distribution

Species-area relationship

Endemics-area relationship

Energy distribution over individuals and species

And also some new metrics:

The distribution of species over genera

The genus-area relationship

Distribution of abundances and metabolic rates over species within a genus with m species.

Genus abundance distribution

Solutions of GSNE:

(Empirical tests now underway)

$$Q(m,n,\varepsilon) = \frac{1}{Z(\lambda_1,\lambda_2,\lambda_3)} e^{-\lambda_1 m} e^{-\lambda_2 mn} e^{-\lambda_3 mn\varepsilon}$$

Master Distribution

$$< m >= \frac{S_0}{G_0} \approx \frac{1}{\lambda_1 \ln(\lambda_1^{-1})} \qquad < n_G >= \frac{N_0}{G_0} \approx \frac{1}{\beta \ln(\beta^{-1})} \qquad \lambda_3 = \frac{G_0}{E_0 - N_0}$$

Determining Lagrange multipliers
($\beta = \lambda_1 + \lambda_2$)

$$\Gamma(m) \approx \frac{e^{-\lambda_1 m}}{m \log(\lambda_1^{-1})}$$

Distribution of Species over Genera

$$\Phi(n) \cong \frac{\lambda_1 \cdot e^{-(\lambda_1+\beta n)}}{n \ln(\beta^{-1})(1 - e^{-(\lambda_1+\beta n)})}$$

Distribution of Individuals over Species

$$\Psi(\varepsilon) \approx \frac{\beta\lambda_3 \cdot \ln(\lambda_1 + \gamma(\varepsilon))}{\gamma^2(\varepsilon)\ln(\beta+\lambda_1)}$$

Distribution of Metabolic Rates over individuals
($\gamma(\varepsilon) = \lambda_2 + \lambda_3\varepsilon$)

$$\Theta(\varepsilon \mid m,n) = \lambda_3 mne^{-\lambda_3 mn(\varepsilon-1)}$$

Distribution of metabolic rates in species with n individuals in a genus with m species

Energy Equivalence:
<metabolic rate> ~ 1/abundance

The SNE theory predicts when it should hold:

The within-species distribution of metabolic rates:

$$\Theta(\varepsilon \mid n) = \frac{R}{\Phi} = \lambda_2 n e^{-\lambda_2 n(\varepsilon - 1)}$$

n = abundance of species

➔ Total energy requirement of a species with abundance n:

$$\int \boldsymbol{n\varepsilon}\ \theta(\varepsilon|n)d\varepsilon = \boldsymbol{n <\varepsilon> = n + 1/\lambda_2}$$

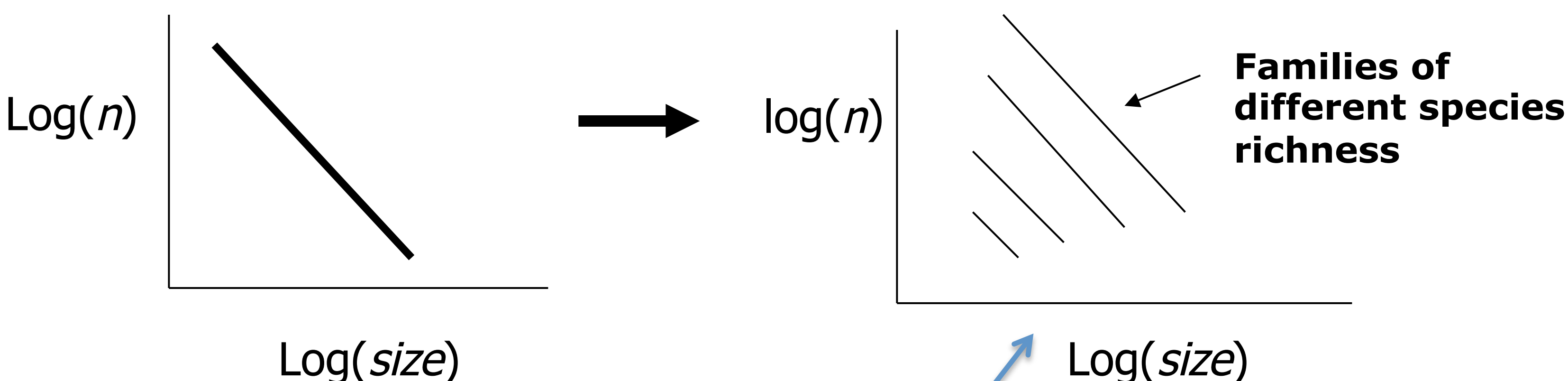
➔ Species obey energy equivalence if:

$$\mathbf{n << 1/\lambda_2 = (E_0 - N_0)/S_0} \,.$$

Including higher taxonomic levels as constraints

If (S,N,E) ⟶ (F,S,N,E) (F = family or other higher order category)

then the predicted size-abundance relationship is modified:



Nature looks more like this

The theory fails to predict patterns in ecosystems undergoing relatively rapid change

1.

Abundance distribution of Rothampsted Moths

Relatively undisturbed fields:
Fisher log series distribution
(predicted by METE)

Fields recently fallowed and in transition:
Lognormal distribution

Kempton and Taylor (1974)

2.

Abundance distribution of trees in Smithsonian tropical forest plots

The most disturbed plot (Barro Colorado Island in Panama) shows the most deviation from METE

3.

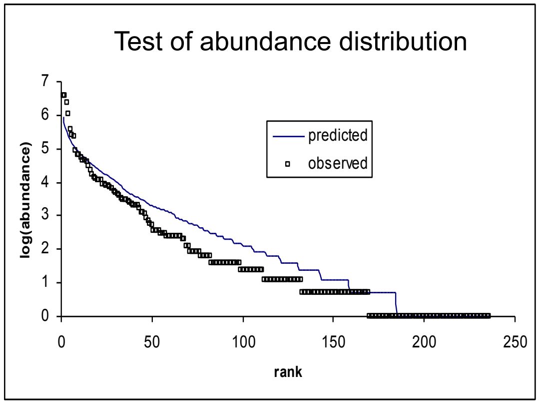
Abundance distributions of Hawaiian Arthropods

sites of different ages and stages of diversification

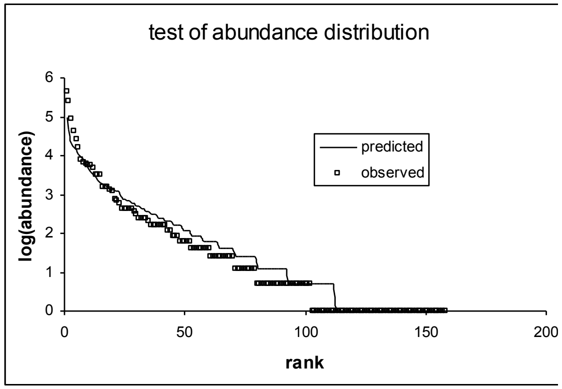
Data from Dan Gruner

Similar pattern of success and failure for body size distributions!

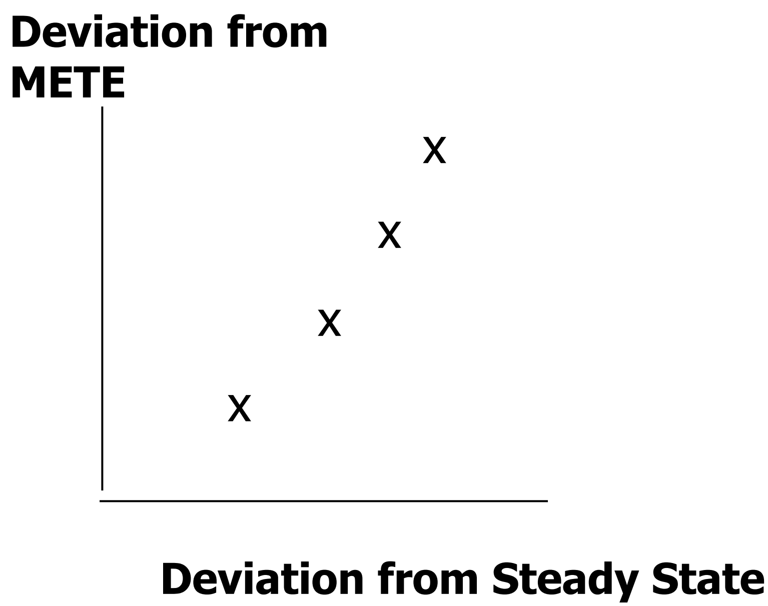
150 y



4 My



Hypothesis:



Unification

time

Heat	⟷	Kinetic Energy
Electricity	⟷	Magnetism
Mass	⟷	Energy
Space	⟷	Time
Information	⟷	Energy
Information theory	⟷	Thermodynamics & Stat. Mech.
Weak force	⟷	Elect. & Magnetism.

It's inconceivable to me that this is where this remarkable story ends

?

today

Unification of all fundamental forces

Information Theory	⟷	Gen. Relativity & Quantum Mech.

(*Verlinde, e.g.,* arXiv:1001.0785v1)

Information Theory

?

"Its from Bits": The laws of nature can be cast in terms of information….only information is truly fundamental.

John Archibald Wheeler, 1990. "Information, physics, quantum: The search for links" in W. Zurek (ed.) *Complexity, Entropy, and the Physics of Information*. Addison-Wesley.

Summary (I)

The MaxEnt principle and specification of a few state variables predicts, with no adjustable parameters,
realistic expressions for:

- species-level spatial abundance distributions
- relative abundance distribution (Fisher log-series), collector's curve
- species-area and endemics-area relationships
- intra-specific and inter-specific metabolic rate distributions
- distribution of linkages across nodes in plant-pollinator & other food webs

Summary (II)

And the theory predicts

- the scale collapse of all species-area relationships onto a universal curve
- species richness at biome scale from small plot data
- the conditions under which energy equivalence should hold

But it appears to poorly predict ecological patterns during periods of rapid change, such as following disturbance. Why?

Can we use this to infer mechanisms that dominate the dynamics?

Thanks:

To my Collaborators:

Erin Conlisk	Adam Smith	Xiao Xiao	Mark Wilber
Justin Kitzes	Andrew Rominger	Ethan White	Chloe Lewis
Erica Newman	Tommaso Zillio	Yu Zhang	Wenyu Zhang

To Other Sources of Data:

J. Green, R. Krishnamani, J. Godinez, W. Kunin, Brian Enquist,
R. Condit, P. Harnik, K. Cherukumilla, E. White, Cory Merow,
D. Gruner, J. Goddard, STRI, David Storch, D. Bartholomew, Yves Basset

To the Funders:

NSF, Miller Foundation, Gordon and Betty Moore Foundation

To my Hosts during the development of METE:

Santa Fe Institute, Rocky Mountain Biological Laboratory, NCEAS, The Chilean Ecological Society, Charles U., U. de Catolica, U. of Padua

& Thank you for listening!

Questions?

What about mechanism? **(Where are the gears?)**

Three responses:

1. The **state variables embody the mechanisms**.

   Success does not imply mechanism does not matter, but rather suggests that the mechanisms incorporated into the values of the state variables suffice; no further mechanistic assumptions are needed.

2. **Mechanism only needed when MaxEnt fails**.

   Analogy with PV = nRT and van der Walls force.

3. **Analogy.** What's the mechanism behind:

   statistical mechanics? or

   quantum mechanics?

Guiding Philosophy:

• MaxEnt is a “null theory”. Just as we learn a lot when a null hypothesis is shown to fail, we can learn a lot when a null theory fails.

• Success does not imply mechanism does not matter! Mechanisms incorporated into the values of the **state variables** suffice; no further mechanistic assumptions are needed.

• Failure of the theory tells us that more mechanistic information than is captured by the state variables is needed to predict patterns in ecology.

A note on alternative measures of entropy

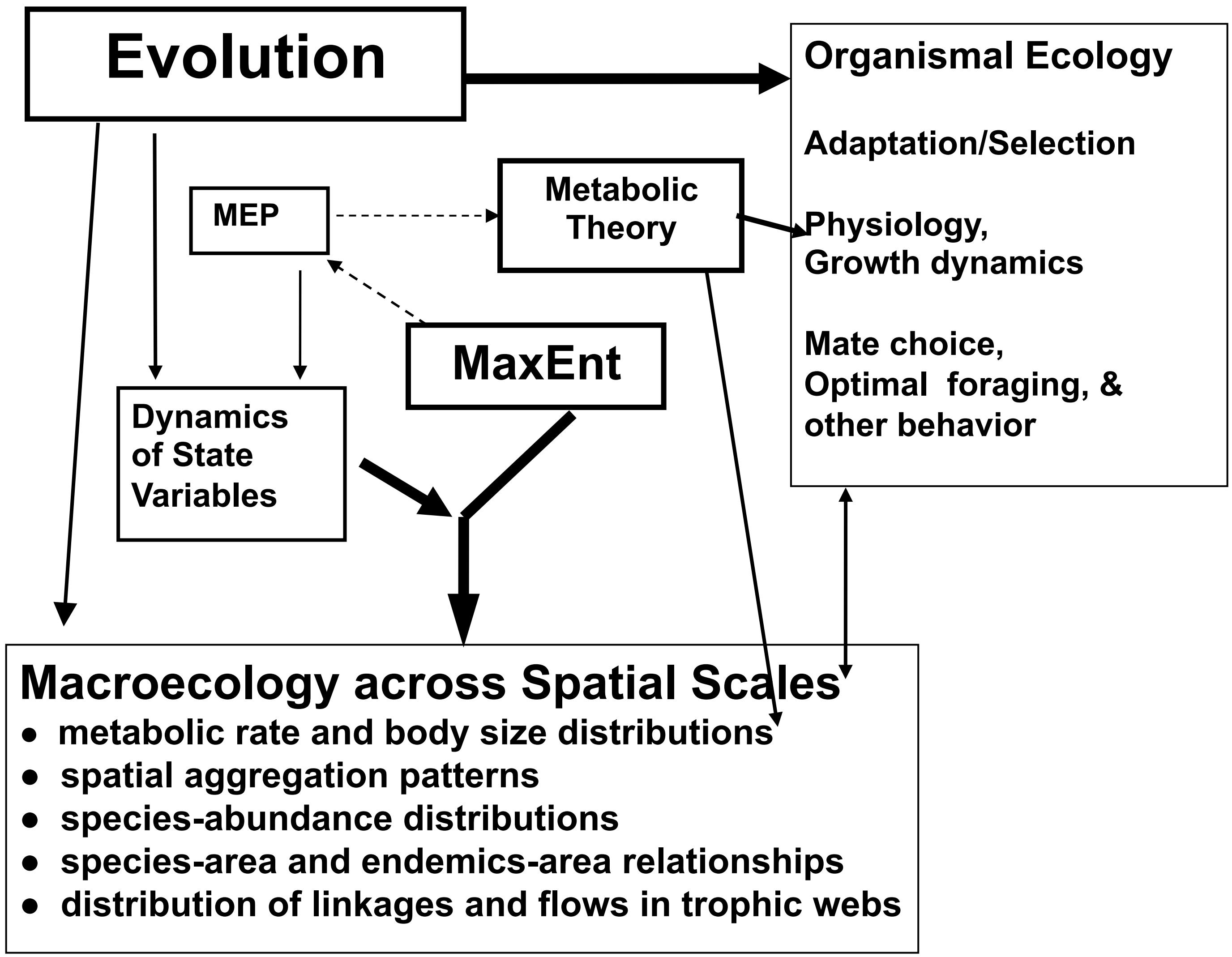
Tsallis entropy: $I_{Tsallis} = (q-1)^{-1} (1 - \Sigma_n [p(n)]^q)$

$\rightarrow I_{shannon}$ as $q \rightarrow 1$

Only empirical information can select the correct q:

The Π distribution describing spatial aggregation only matches data for $q \sim 1$

Unification: why let physicists have all the fun?



MILESTONES IN THE DEVELOPMENT OF INFORMATION THEORY

Laplace: Rule of Succession

Shannon: Information entropy of a message

Landauer: 1 bit = kT

Jaynes: Objective Bayes, MaxEnt

Wheeler: "Its from Bits": The suggestion that the laws of physics can be cast in terms of information….only information is truly fundamental.

Dice problem

Suppose all you know is that a die that has been flipped 10,000 times had a mean score of 3.5.

What should you assume for $P(n)$: $n = 1, \ldots, 6$?

We know MaxEnt insures that P will be of the form $P(n) = e^{-\lambda n}/Z = x^n/Z$

$(x = e^{-\lambda})$

The Lagrange multiplier gives two equations to solve:

$$< n > = \frac{\sum_1^6 n x^n}{\sum_1^6 x^n} = 3.5 \qquad \sum_1^6 x^n / Z = 1$$

The solution is: $x = 1$ (or $\lambda=0$), $Z = 6$.

$P(n) = 1/6$ for all n

We have just cut butter with a chainsaw….

A harder problem: an unfair die

Suppose you have a 3-faced die: n = 1, 2, 3

and that when flipped 10,000 times the mean is had a mean score of 1.5, not 2.

What should you infer for P(n): n = 1, 2, 3

Again P must be of the form $P(n) = e^{-\lambda n}/Z = x^n/Z$.

The Lagrange multiplier calculation now gives:

$$< n > = \frac{\sum_1^3 n x^n}{\sum_1^3 x^n} = 1.5 \qquad \sum_1^3 x^n / Z = 1$$

The solution is $x = (-1 \pm \sqrt{13})/6$; $Z = 0.705$

Only + gives real-valued P(n)

P(1) = 0.616; P(2) = 0.268; P(3) = 0.116

Homework: do the calculations and derive the results above.

Suppose all you know is that P(n’) = a.

Let f(n) = $\delta_{n,n'}$. This is the Kronecker delta: = 1 if n = n’ ; = 0 if n ≠ n’ .

$$\sum_{n=0}^{N} P(n)\delta_{n,n'} = a$$ i.e., $f(n) = \delta_{n,n'}$, $F = a$

$$P(n) = \frac{e^{-\lambda\delta_{n,n'}}}{Z}$$

where:

$$Z = \sum_{n=0}^{N} e^{-\lambda\delta_{n,n'}} = \sum_{n=0}^{N} [1 - \lambda\delta_{n,n'} + \frac{\lambda^2\delta_{n,n'}{}^2}{2} - \dots]$$
$$= N + 1 - \lambda + \lambda^2/2 - \dots = N + e^{-\lambda}$$

$$-\frac{\partial Z}{Z\partial\lambda} = a \quad \rightarrow \quad e^{-\lambda} = aZ = a(N + e^{-\lambda}) \quad \rightarrow \quad e^{-\lambda} = \frac{aN}{1-a}$$

$$\boxed{P(n) = \frac{1-a}{N}\left(\frac{aN}{1-a}\right)^{\delta_{n,n'}}}$$

Homework: work out the answer for two point constraints or for 1 point constraint + knowledge of <n>

MaxEnt and Environmental Envelopes

Environmental variable, *T*	Species status
1	0
2	1
3	-
6	1
4	-
5	0
3	-
5	0
2	-
1	1

We want to infer the values of the missing (1,0) entries

Step 1: $P(T|1) = ?$
Step 2: Use Bayes to derive $P(1|T)$

Like a 6-sided die problem: <T> given a 1 = (2+6+1)/3= 3

Lagrange multiplier: $x+2x^2+3x^3+4x^4+5x^5+6x^6 = 3(x+x^2+x^3+x^4+x^5+x^6)$

implies $x = \exp(-\lambda) = 0.8398$; $Z = 3.4033$

$P(T|1) = (0.8398)^T/3.4033$

$P(1|T)= P(T|1)P(1)/P(T)$. $P(1) =1/2$, $P(T)$: either take each to be equally likely, so $P(T) = 1/6$ for all T, Then, e.g., $P(1|3) = [(0.8398)^3/3.4033][(1/2)/(1/6)]=0.52$

Or use the environmental data to get $P(T)$.

Suppose we want to know the number of species of beetles or spiders or orchids or trees in Amazonia.

Available data might consist of presence-absence information in a large number (perhaps 100) of small plots or fumigated trees scattered randomly throughout Amazonia.

If we knew the form of the species area relationship across the entire scale range from plot to Amazonia we'd be done. For, example, suppose $S(A) = cA^z$. Then

$$S(A_0) = S(A)\,(A_0/A)^z$$

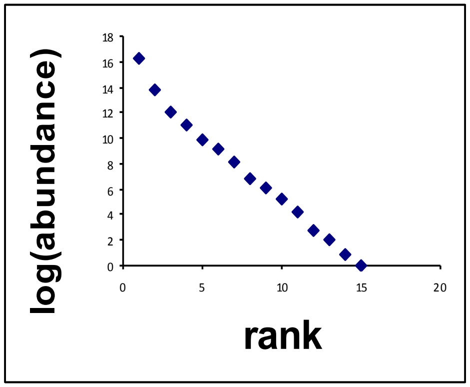
So letting A be the small plot area, we can scale up.

We need to know the form of the SAR across that huge scale range.

Some Prevalent Scaling Patterns in Macroecology

The distribution of abundances across species

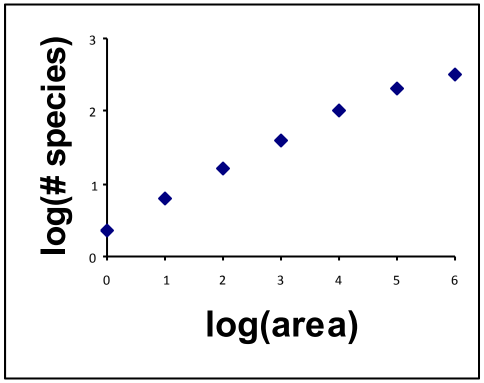
(*"Fisher log-series, Lognormal, ???*)



Species-area relationship

(dependence of # species on area sampled is <u>sometimes</u>

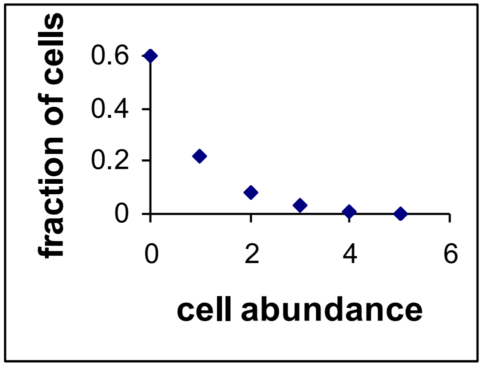
taken to be a power law, but often curvature on log-log plot)



The distribution of individuals (within species)

in cells of arbitrary scale: i.e., $\Pi(n \text{ in } A|n_0,A_0)$

(*often exponential decrease observed at many scales;*



Why do we care about patterns and metrics in ecology?

1. **Extinction rates under habitat loss.**

   25% of Amazon rain forest has been cut.
   How many species lost?

2. **Scaling up biodiversity.**

   How many species of arboreal beetles
   in all of the Amazonian rain forest?

3. **Inferring process from pattern**
   (analogy: *Brahe -> Kepler -> Newton*)

Data $\longrightarrow$ **Pattern** $\longrightarrow$ **Force Law**

Access to plant census data from a serpentine grassland.

The following web site contains a spatially-explicit vegetation data set:

http://conium.org/~hartelab/MaxEnt.html

The census was carried out over the spring and summer of 1998 by Jessica Green on a 64 m^2 plot at the University of California's McLaughlin Reserve in xx county, CA. The plot was gridded to a smallest cell size of ¼ m^2 and in each cell the abundance of every plant species found there was recorded.

The columns are plant species, with each species given a code name explained below the table of data. There are 256 rows of data, with each row corresponding to one of the ¼ m^2 cells. If the plot is viewed as a matrix, then the first row of data in the spread sheet corresponds to the upper left cell (matrix element a_{11}). The second row of data is the matrix element a_{12}, or in other words the cell just to the right of a_{11}. The 17th row of data then corresponds to the plot matrix element a_{21}, and the very last row of data is the lower right cell, $a_{16,16}$. The actual data entries are the abundances of the species in each cell.

The data may be used by readers for any purpose, but any publication that includes use of the data should reference the data set to:

Green, J., Harte, J., and Ostling, A., (2003). Species richness, endemism, and abundance patterns: tests of two fractal models in a serpentine grassland. *Ecology Letters* 6, 919-928. Moreover, the Acknowledgments should include a thanks to Jessica Green for use of the data.

Howmwork:

1. Plot the SAR on a log(S) vs. log(area) and as a slope versus ln(N/S) graph

2. Plot the SAD as a rank abundance graph.

Streamlining data analysis and theory testing

macroeco

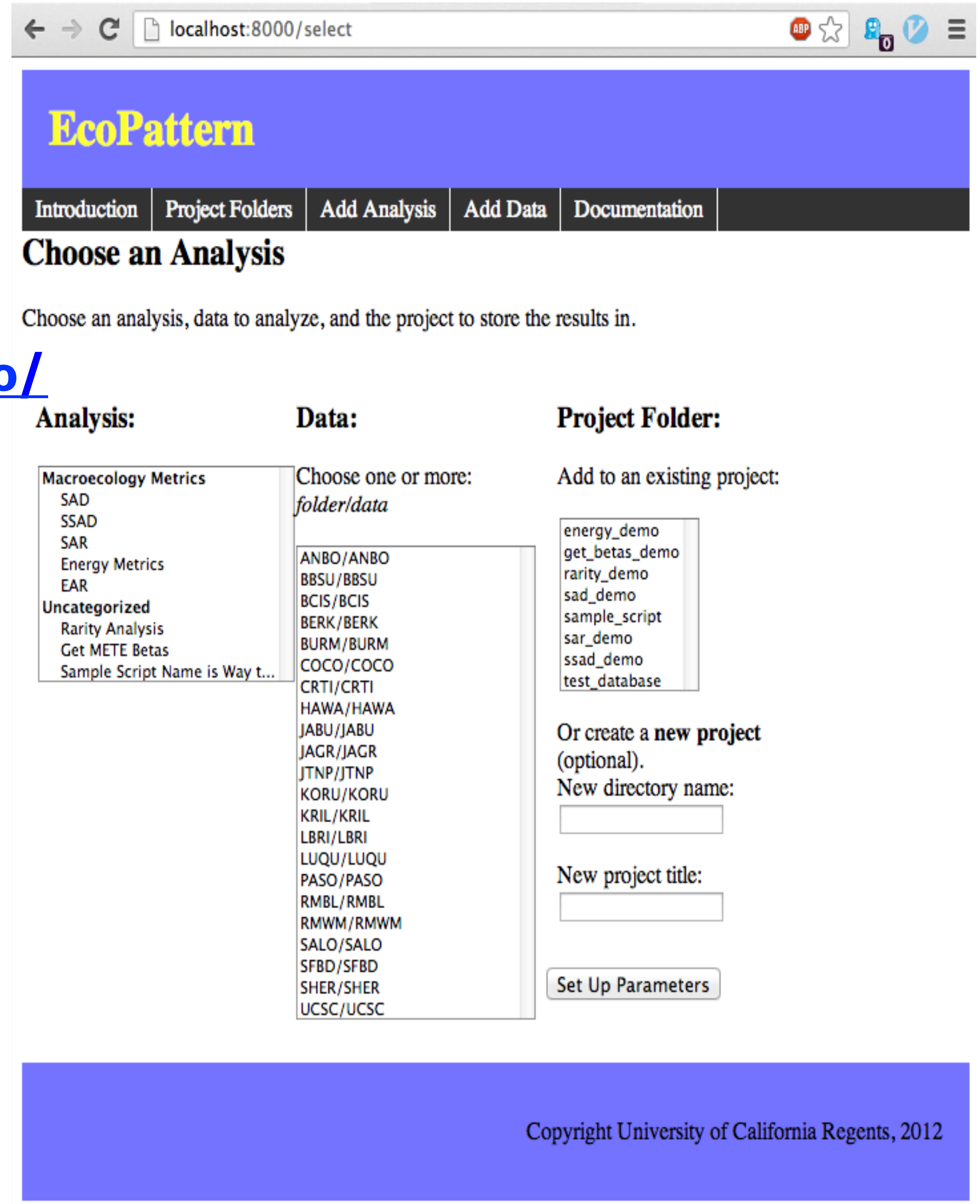
Python package for ecological data analysis and theory comparison

http://jkitzes.github.io/macroeco/

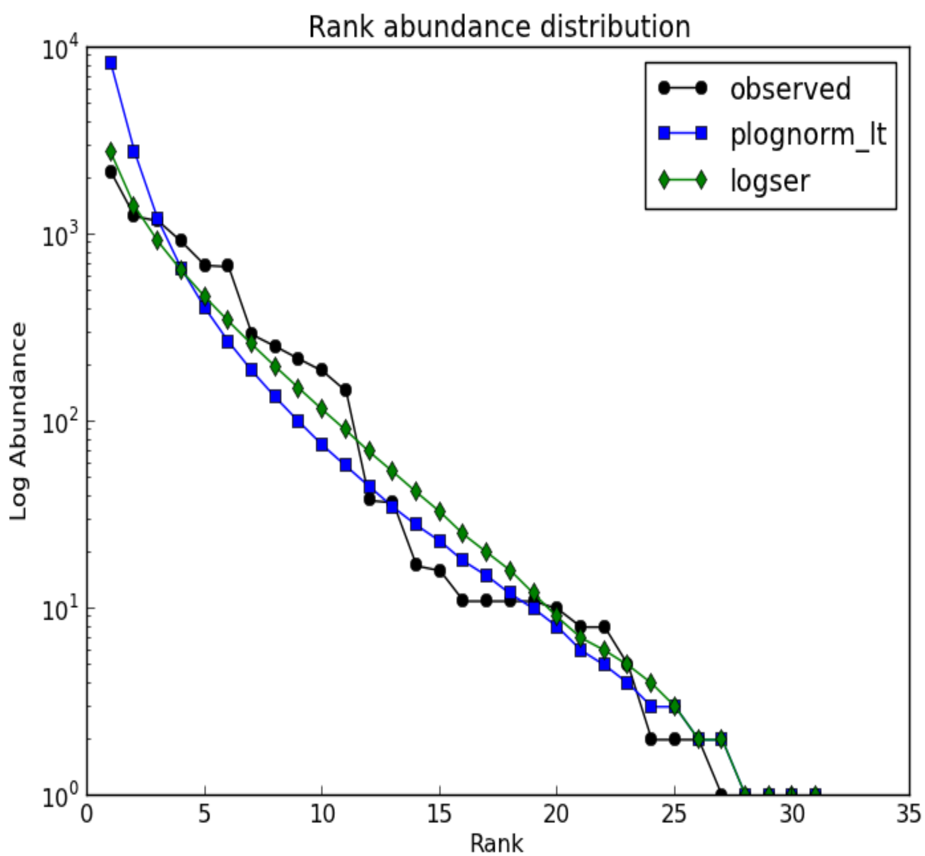
ecopattern

Graphical user interface and data management system for using macroeco

Justin Kitzes
Chloe Lewis
Mark Wilber



Example of "push button" output for Gilbert forest census, Santa Cruz



Criteria for plot: {}

```
CRITERIA: {}

EMPIRICAL VALUES:
Species = 31
Total Individuals = 8180.0
Observed Nmax = 2162.0
Observed Rarity = {'<=10': 12}

PREDICTED DISTRIBUTION : plognorm_lt
Species = 31
Total Individuals = 14264.0
AIC = 344.411070479
Delta_AIC = 6.13188971425
AIC_weight = 0.0445340587232
Number of Parameters = 2
Predicted Nmax = 8181.0
Predicted Rarity = {'<=10': 13}
Other Variables = {}

PREDICTED DISTRIBUTION : logser
Species = 31
Total Individuals = 7660.0
AIC = 338.279180765
Delta_AIC = 0.0
AIC_weight = 0.955465941277
Number of Parameters = 1
Predicted Nmax = 2739.0
Predicted Rarity = {'<=10': 12}
Other Variables = {'p': 0.9995019094596226}
```

Unification

time

Heat ↔ Kinetic Energy

Electricity ↔ Magnetism

Mass ↔ Energy

Space ↔ Time

Information ↔ Energy

Information theory ↔ Thermodynamics & Stat. Mech.

Weak force ↔ Elect. & Magnetism.

It's inconceivable to me that this is where this remarkable story ends

?

today

Unification of all fundamental forces

Information Theory ↔ Gen. Relativity & Quantum Mech.

(Verlinde, e.g., arXiv:1001.0785v1)

?

Information Theory