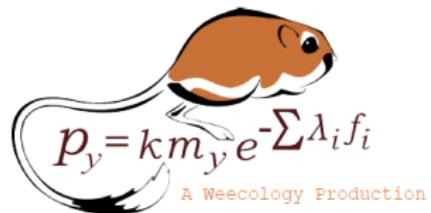


# A DATA-INTENSIVE ASSESSMENT OF THE SPECIES-ABUNDANCE DISTRIBUTION.

Elita Baldridge



# OPEN SCIENCE

- Code:
  - [github.com/embaldridge](https://github.com/embaldridge)
  - [github.com/weecology](https://github.com/weecology)
- Data: [figshare.com](https://figshare.com)
- Twitter: @elitabaldridge



**figshare**  
credit for all your research



# Feel free to:



Copy, share, adapt, or re-mix;



Photograph, film, or broadcast;



Blog, live-blog, or post video of;

# Provided that:



You attribute the work to its author and respect the rights and licenses associated with its components.

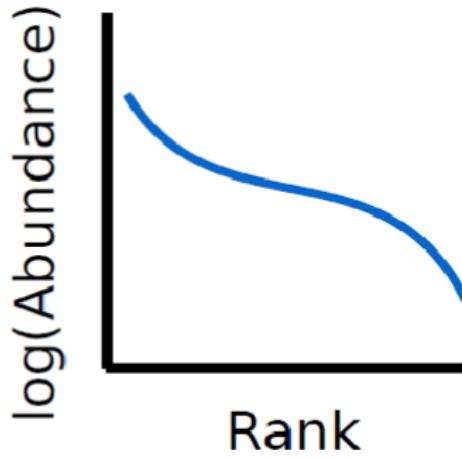
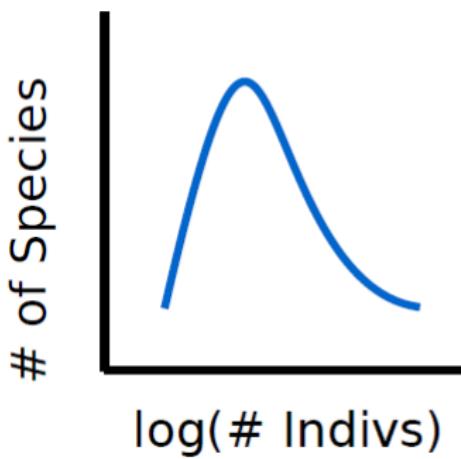
## COMMONNESS & RARITY

"Who can explain why one species ranges widely, and is very numerous, and why another allied species has a narrow range and is rare? Yet these relations are of the highest important, for they determine the present welfare and, as I believe, the future success and modification of every inhabitant of this world."

Darwin, 1859.

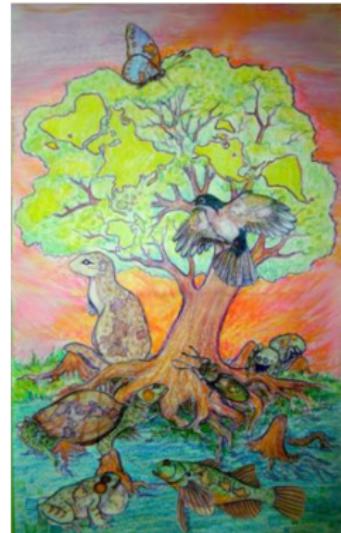
## Species abundance distribution

- Describes the distribution of commonness & rarity of species.
- Exhibits a hollow curve distribution.



# MACROECOLOGY

One approach to studying ecological patterns and processes.



MACROECOLOGY

Pattern



Process



Prediction

# MACROECOLOGY

## Challenges of macroecology

- Studies performed with a limited number of large datasets.
- Lack of identification of pattern generating mechanisms.

# MACROECOLOGY

## Best practice recommendations

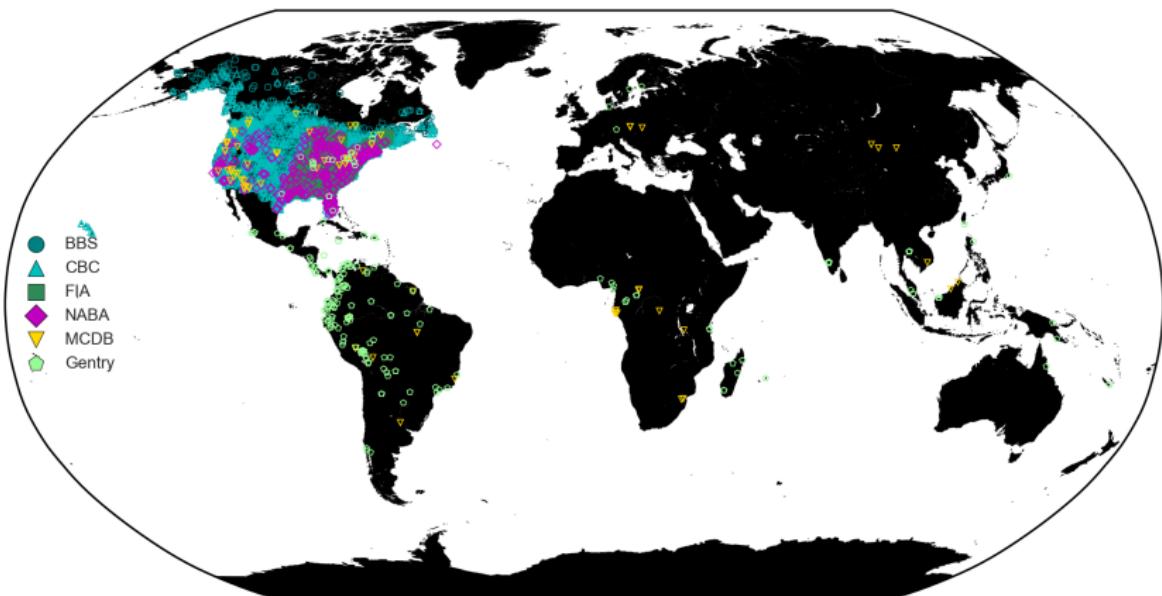
- Test patterns with multiple taxonomic groups/ecosystems.
- Simultaneous testing of competing models and model predictions with a consistent statistical approach.

# THE RULES OF ECOINFORMATICS

Garbage in, garbage out.

- All data are good, not all data are appropriate.
- Fit the data to the question.

# DATA



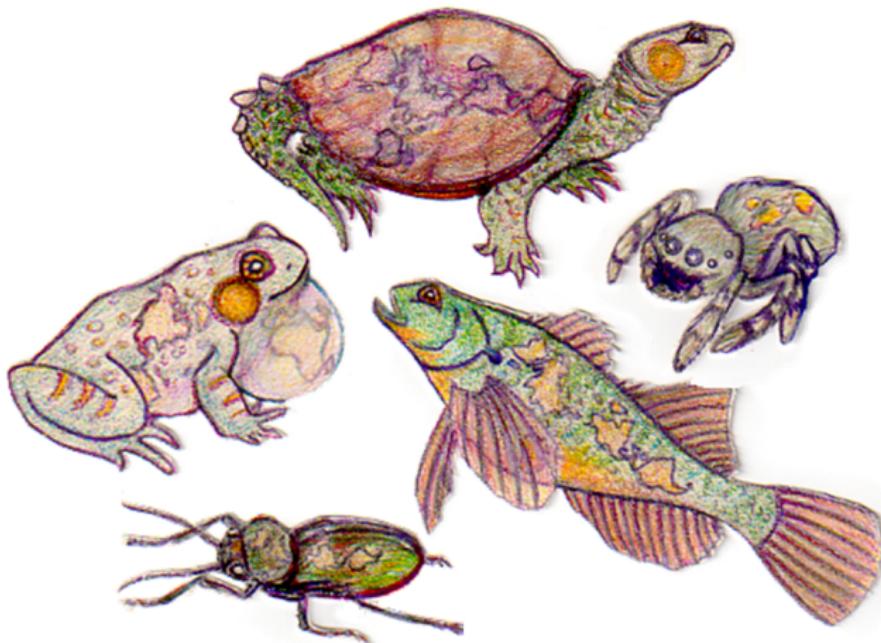
# DATA

## Major macroecological datasets

- Largely terrestrial
- Largely North American
- Many publicly available, some not.

Lots of data in the literature.

# DATA



# DATA

Inclusion criteria:

- Quantitative abundances
- Animals
- Complete sampling
- Must not be heavily summarized or processed
- High degree of taxonomic resolution
- Observational

# DATA WRANGLING

1.pdf - Adobe Reader  
File Edit View Window Help  
Open Tools Fill & Sign  
p. 14 (4 of 10) 150% Tools View  
14 JOHN F. CAVITT

TABLE 1. Number of individuals and relative abundance (#/100 Trap array days) by year and site for the 10 species of snakes and three species of lizards captured (\* indicates focal species).

Site	A			B			C			D		
	1994	1995	1996	1995	1996	1995	1996	1995	1996	1995	1996	Total
<b>Snakes</b>	4	9	9	9	9	4	4	4	4	4	4	
* <i>Crotalus cerastes</i>	33	44	52	58	20	4	31	15	28	28	28	
* <i>Thamnophis sirtalis</i>	9.37	5.82	6.64	7.67	2.55	1.82	7.75	7.81	7.00			
<i>Elaphe emoryi</i>	5	12	9	15	10	21	10	26	7	115		
<i>Elaphe emoryi</i>	1.42	1.59	1.15	1.98	1.28	9.55	2.50	13.54	1.75			
<i>Lampropeltis getula</i>	2	13	3	3	7	5	3	8	4	48		
<i>Lampropeltis getula</i>	0.57	1.72	0.38	0.40	0.89	2.27	0.75	4.17	1.0			
<i>Lampropeltis triangulum</i>	5	5	6	7	7	1	1	1	5	38		
<i>Lampropeltis triangulum</i>	1.42	0.66	0.77	0.93	0.89	0.45	0.25	0.52	1.25			
<i>Pituophis catenifer</i>	3	1	3	7	4	—	1	3	4	26		
<i>Pituophis catenifer</i>	0.85	0.13	0.38	0.93	0.51	—	0.25	1.56	1.0			
<i>Elaphe obsoleta</i>	—	—	2	3	—	2	2	5	12	26		
<i>Elaphe obsoleta</i>	—	—	0.26	0.40	—	0.91	0.5	2.60	3.0			
<i>Tropidoclonion lineatum</i>	4	1	1	—	2	—	—	—	—	8		
<i>Tropidoclonion lineatum</i>	1.14	0.13	0.13	—	0.26	—	—	—	—			
<i>Lampropeltis calligaster</i>	2	—	—	—	—	—	—	—	—	2		
<i>Lampropeltis calligaster</i>	0.57	—	—	—	—	—	—	—	—			
<i>Storeria dekayi</i>	—	—	1	—	—	—	—	—	—	1		
<i>Storeria dekayi</i>	—	—	0.13	—	—	—	—	—	—	0.52		
<b>Lizards</b>	—	—	—	—	—	—	—	—	—	—		
* <i>Ophisaurus attenuatus</i>	17	22	10	30	10	2	—	—	—	91		
* <i>Ophisaurus attenuatus</i>	4.83	2.91	1.28	3.97	1.28	0.91	—	—	—			
<i>Eumeces obsoletus</i>	1	2	—	6	1	—	1	—	—	11		
<i>Eumeces septentrionalis</i>	0.28	0.26	—	0.79	0.13	—	0.25	—	—	6		
<i>Eumeces septentrionalis</i>	—	1	—	5	—	—	—	—	—			
<i>Eumeces septentrionalis</i>	—	0.13	—	0.66	—	—	—	—	—			

Species\_abundances.csv - Microsoft Excel  
Home Insert Page Layout Formulas Data Review View  
Clipboard A1 Class  
A B C D E F G  
1 Class Family Genus Species Relative\_Abundance Site\_ID  
2 Reptilia Pituophis catenifer 0 0  
3 Reptilia Lampropeltis calligaster 0 0  
4 Reptilia Storeria dekayi 0 0  
5 Reptilia Eumeces septentrionalis 0 0  
6 Reptilia Eumeces obsoletus 0.28 1  
7 Reptilia Elaphe emoryi 1.72 2  
8 Reptilia Tropidoclonion lineatum 0.57 2  
9 Reptilia Lampropeltis triangulum 0.85 3  
10 Reptilia Elaphe obsoleta 1.14 4  
11 Reptilia Thamnophis sirtalis 1.42 5  
12 Reptilia Lampropeltis getula 0.66 5  
13 Reptilia Ophisaurus attenuatus 4.83 17  
14 Reptilia Coluber constrictor 9.37 33  
15 Reptilia Pituophis catenifer 0 0  
16 Reptilia Tropidoclonion lineatum 0 0  
17 Reptilia Lampropeltis calligaster 0 0  
18 Reptilia Storeria dekayi 0 0  
19 Reptilia Lampropeltis triangulum 0.13 1  
20 Reptilia Elaphe obsoleta 0.13 1  
21 Reptilia Eumeces septentrionalis 0.13 1  
22 Reptilia Eumeces obsoletus 0.26 2  
23 Reptilia Lampropeltis getula 0.77 5  
24 Reptilia Thamnophis sirtalis 1.59 12  
25 Reptilia Elaphe emoryi 0.38 13  
26 Reptilia Ophisaurus attenuatus 2.91 22  
27 Reptilia Coluber constrictor 5.82 44  
28 Reptilia Tropidoclonion lineatum 0 0  
29 Reptilia Storeria dekayi 0 0  
30 Reptilia Eumeces obsoletus 0 0  
31 Reptilia Eumeces septentrionalis 0 0  
32 Reptilia Elaphe obsoleta 0.13 1  
Reptilia, Pituophis, catenifer, 0, 0, 1, 1  
Reptilia, Storeria, dekayi, 0, 0, 1, 1  
Reptilia, Eumeces, septentrionalis, 0, 0, 1, 1  
Reptilia, Eumeces, obsoletus, 0, 2, 1, 1, 1  
Reptilia, Elaphe, emoryi, 1, 72, 2, 1, 1  
Reptilia, Tropidoclonion, lineatum, 0, 57, 2, 1, 1  
Reptilia, Lampropeltis, triangulum, 0, 85, 3, 1, 1  
Reptilia, Elaphe, obsoleta, 1, 42, 1, 1  
Reptilia, Lampropeltis, getula, 0.66, 5, 1, 1  
Reptilia, Ophisaurus, attenuatus, 1.83, 17, 1, 1  
Reptilia, Coluber, constrictor, 9.37, 33, 1, 1  
Reptilia, Pituophis, catenifer, 0, 0, 2, 1  
Reptilia, Tropidoclonion, lineatum, 0, 0, 2, 1  
Reptilia, Lampropeltis, calligaster, 0, 0, 2, 1  
Reptilia, Storeria, dekayi, 0, 0, 2, 1  
Reptilia, Lampropeltis, triangulum, 0.13, 1, 2, 1  
Reptilia, Elaphe, obsoleta, 0, 13, 1, 2, 1  
Reptilia, Eumeces, septentrionalis, 0.13, 1, 2, 1  
Reptilia, Eumeces, obsoletus, 0, 26, 2, 2, 1  
Reptilia, Lampropeltis, getula, 0.77, 5, 2, 1, 1  
Reptilia, Thamnophis, sirtalis, 1.59, 12, 2, 1, 1  
Reptilia, Elaphe, emoryi, 0.38, 13, 2, 1  
Reptilia, Ophisaurus, attenuatus, 2.91, 22, 2, 1, 1  
Reptilia, Coluber, constrictor, 5.82, 44, 2, 1  
Reptilia, Tropidoclonion, lineatum, 0, 0, 3, 1  
Reptilia, Storeria, dekayi, 0, 0, 3, 1  
Reptilia, Eumeces, obsoletus, 0, 0, 3, 1  
Reptilia, Eumeces, septentrionalis, 0, 0, 3, 1  
Reptilia, Elaphe, obsoleta, 0, 13, 1, 3, 1  
Reptilia, Lampropeltis, calligaster, 0, 13, 1, 3, 1  
Reptilia, Pituophis, catenifer, 0, 26, 2, 3, 1  
Reptilia, Elaphe, emoryi, 0, 43, 3, 1  
Reptilia, Lampropeltis, triangulum, 0, 38, 3, 3, 1  
Reptilia, Lampropeltis, getula, 0, 93, 6, 3, 1  
Reptilia, Thamnophis, sirtalis, 1, 19, 9, 3, 1  
Reptilia, Ophisaurus, attenuatus, 1, 28, 10, 3, 1  
Reptilia, Coluber, constrictor, 0, 64, 52, 3, 1  
Reptilia, Elaphe, obsoleta, 0, 0, 4, 1

# DATA

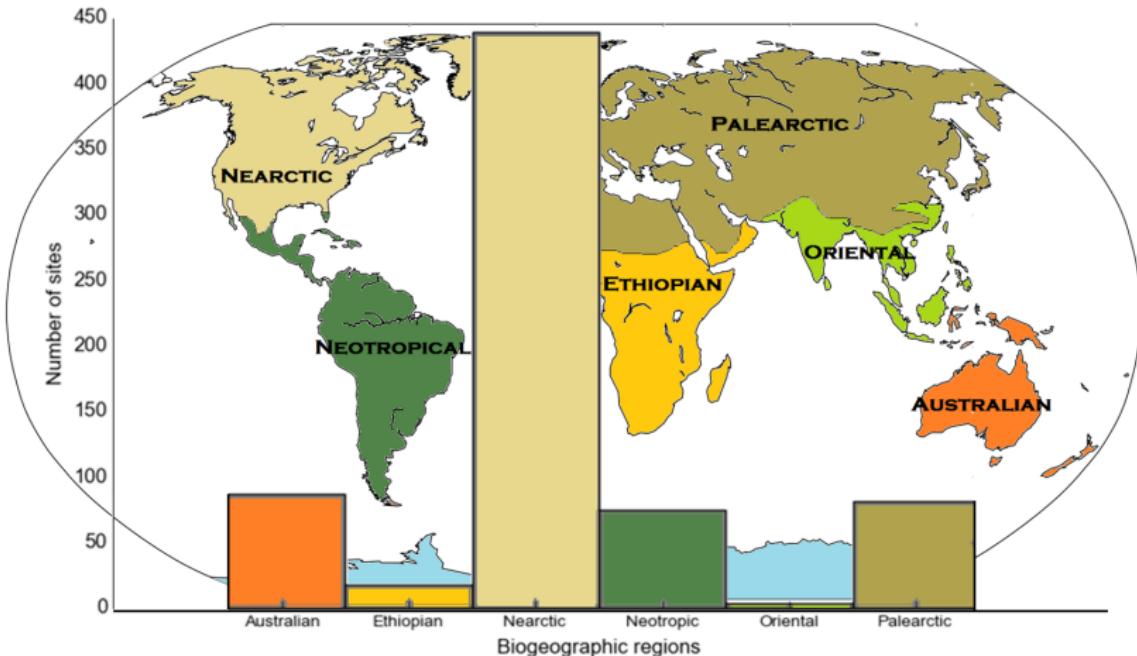
## Variables collected

---

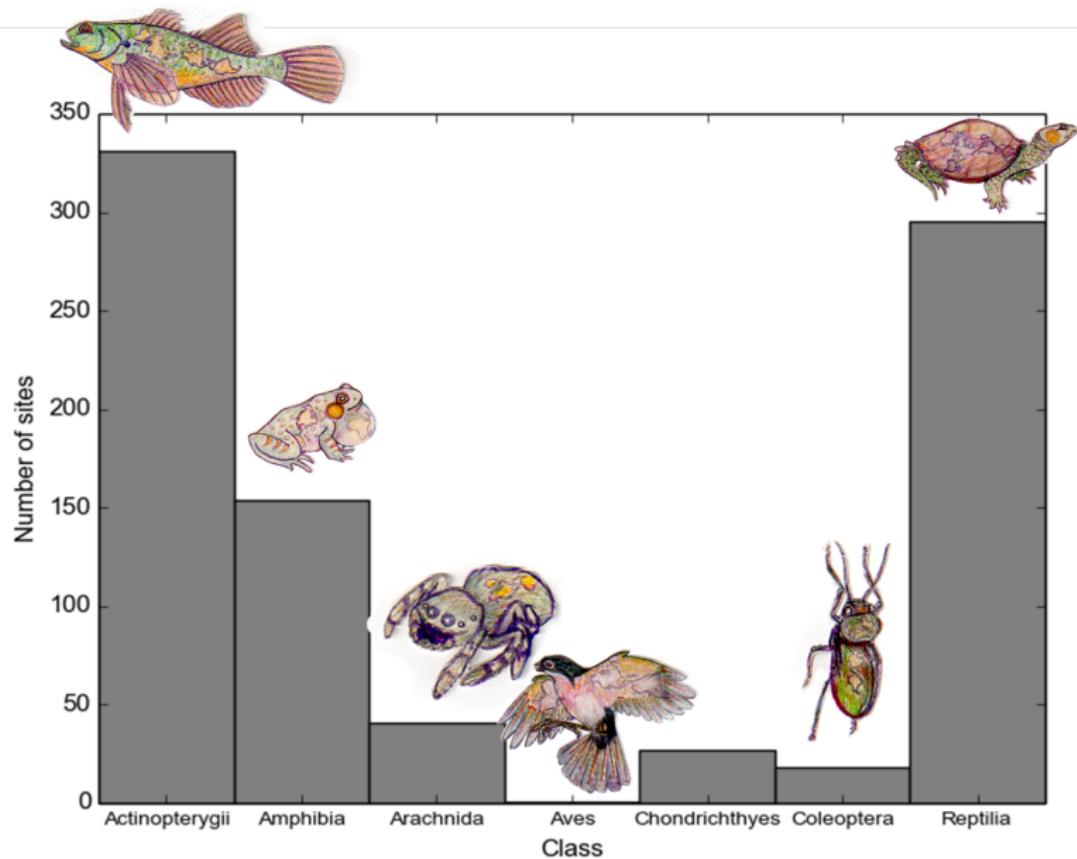
Class  
Family  
Genus  
Species (Specific epithet)  
Relative abundance  
Abundance  
Collection Year, starting  
Collection Year, ending  
Site Name  
Biogeographic region  
Site notes

TABLE : List of variables collected.

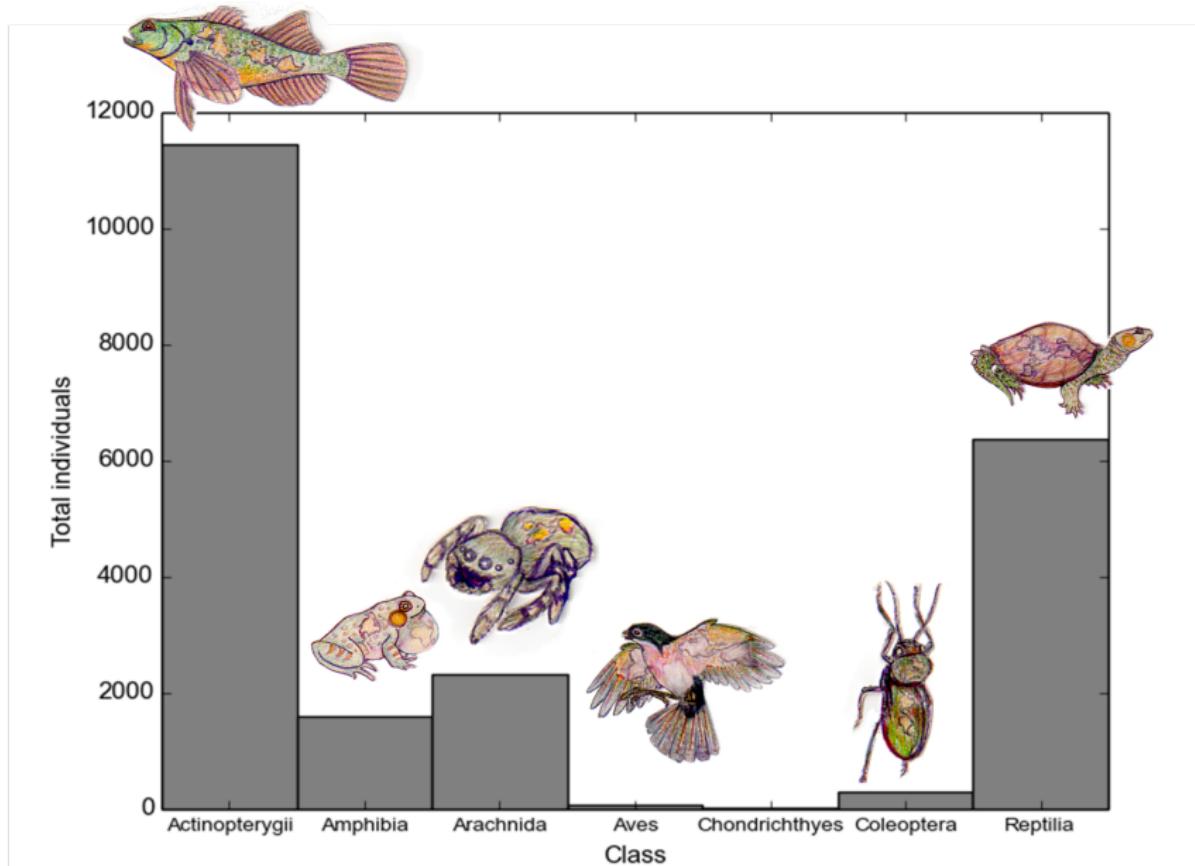
# DATA



# DATA



# DATA



## DATA AVAILABILITY

Public & open access through figshare.  
EcoData Retriever importable.

(<http://figshare.com>)

(<http://www.ecodataretriever.org>)

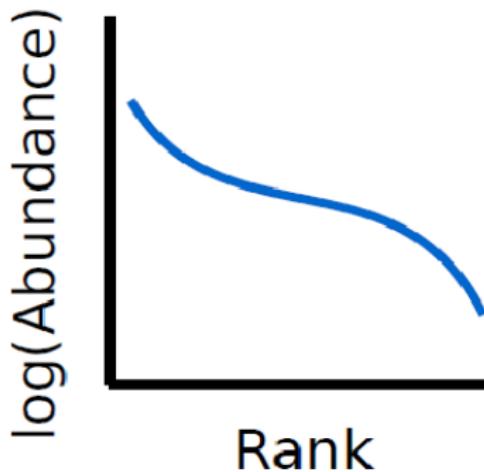
sad\_data =

ecoretriever::fetch('MiscAbundanceDB')



## COMMONNESS & RARITY

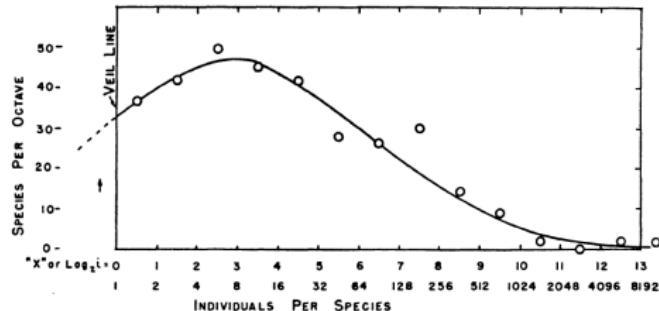
The species abundance distribution:



Many models of the species abundance distribution (SAD).

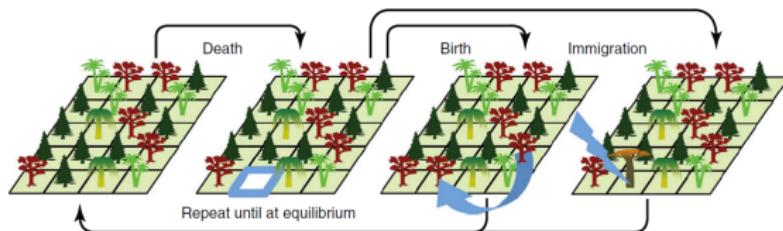
# SAD MODELS

## Statistical description



Preston 1962a.

## Process-based



Rosindell et al. 2011.

## SAD COMPARISONS

Most comparisons of the different models:

- Use only a small subset of available models (typically two).
- Focus on a single ecosystem or taxonomic group
- Fail to use the most appropriate statistical methods.

# SAD COMPARISONS

Selected five models from four classes for comparison.

Model class	Form of the distribution
Purely statistical	Logseries, Poisson lognormal
Branching process	Zipf
Population dynamics	Negative binomial
Niche partitioning	Geometric

TABLE : After B.J. McGill et al. 2007.

# SAD COMPARISONS

Analysis:

- Model fitting with maximum likelihood estimation.

$$l_x(\theta) = \log h(x) + \log f_\theta(x)$$

# SAD COMPARISONS

## Analysis:

- Likelihood based model selection to compare the fits of the different models.
  - Assess the fit of the model without corrections for parameter number or similarity to other models.
  - How well does the model describe the data?

# SAD COMPARISONS

## Analysis:

- Model comparison with corrected Aikaike Information Criterion (AICc) weights.
  - Assess the fit of the model correcting for parameter number and small sample size.
  - How well does the model describe the data relative to the number of parameters?

# SAD COMPARISONS

Computational tools:

- Model fitting, log-likelihood, & AICc: macroecotools Python package.  
(<https://github.com/weecology/macroeccotools>)
- All code & majority of data are publicly available.  
(<https://github.com/weecology/sad-comparison>)



OPEN SCIENCE

More reproducible.  
More accessible.

*Better science.*

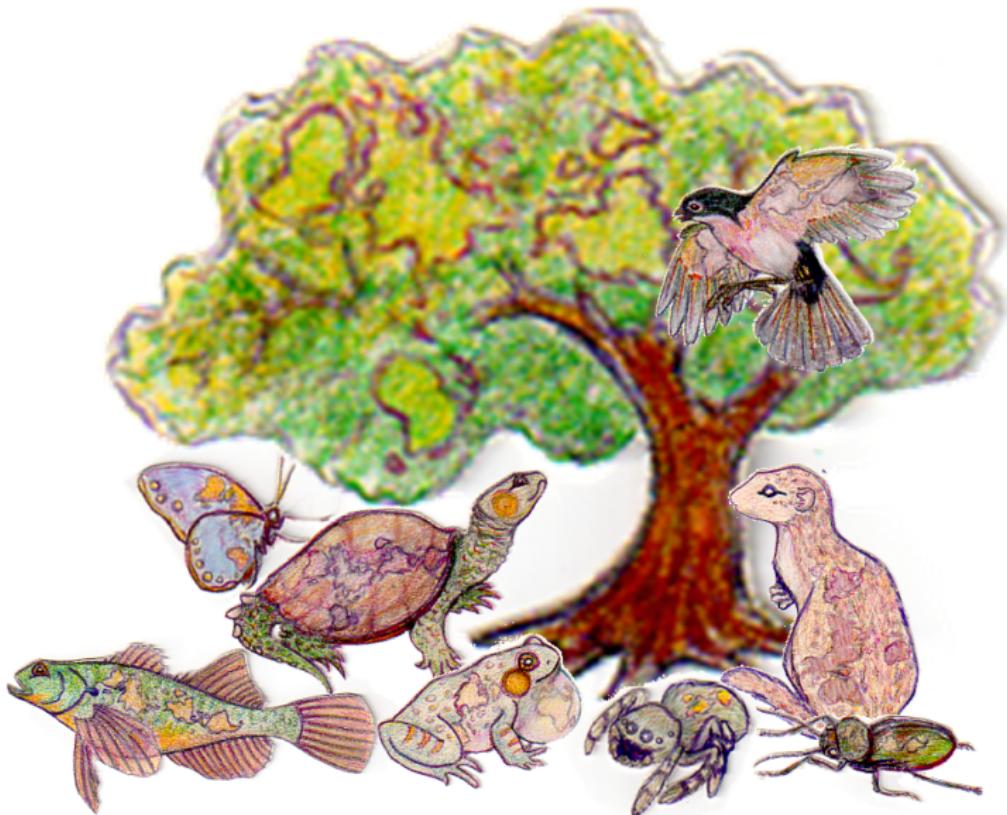


# SAD COMPARISONS

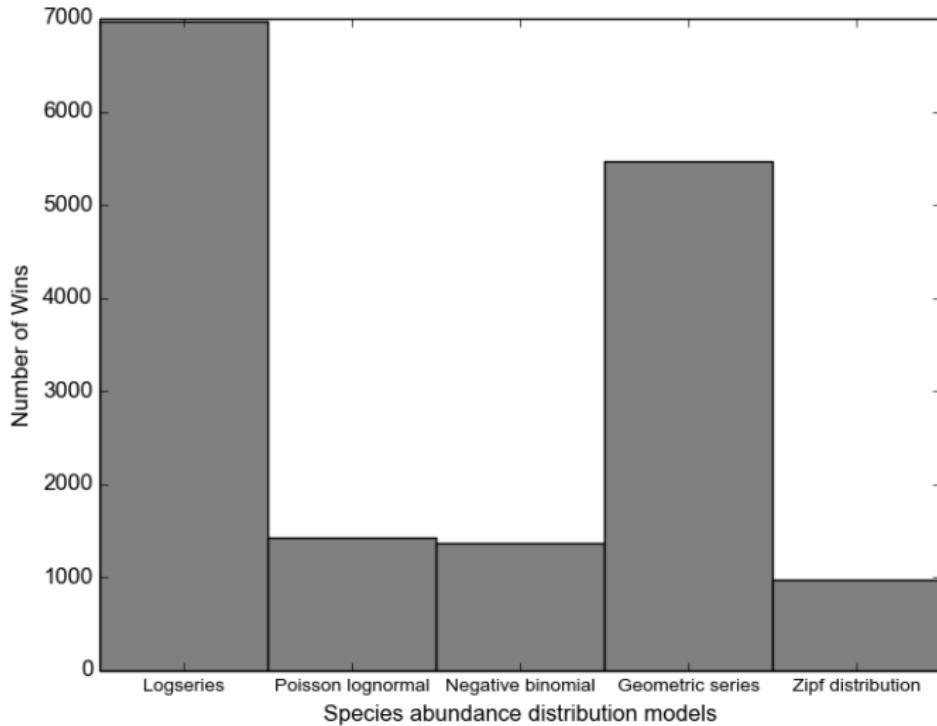
Dataset	Dataset code	Availability	Sites
Gentry's Forest Transects	Gentry	Public	10355
Breeding Bird Survey	BBS	Public	2769
Christmas Bird Count	CBC	Private	1999
Forest Inventory Analysis	FIA	Public	220
N. American Butterfly Count	NABA	Private	400
Actinopterygii, compiled	Actinopterygii	Public	161
Reptilia, compiled	Reptilia	Public	138
Mammal Community Database	MCDB	Public	103
Amphibia, compiled	Amphibia	Public	43
Arachnida, compiled	Arachnida	Public	25
Coleoptera, compiled	Coleoptera	Public	5

TABLE : Datasets used for species-abundance distribution comparisons.  
Datasets marked as Private obtained through data requests to the providers with  
Memorandums of Understanding.

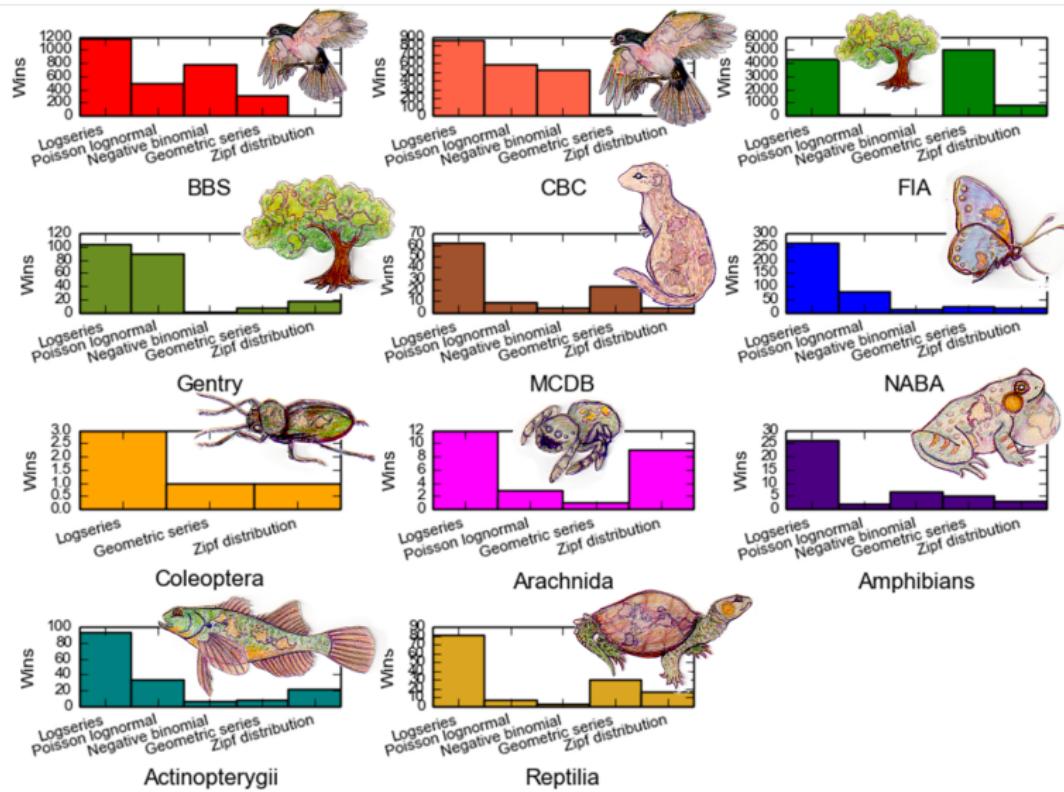
# DATA



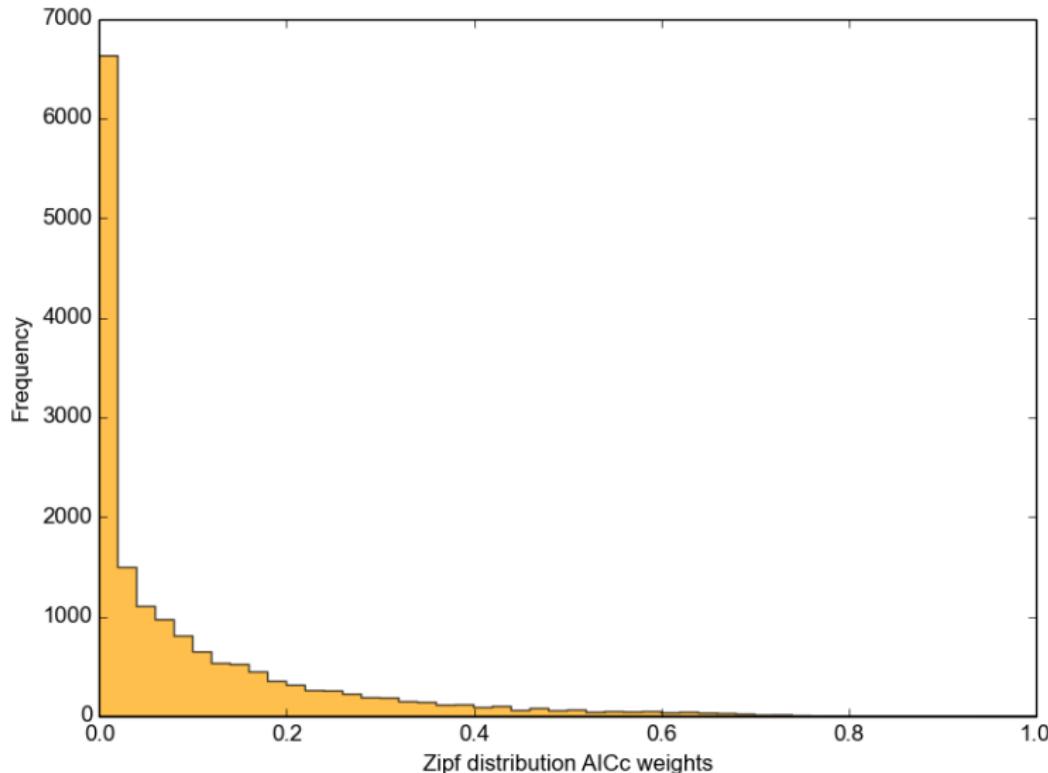
# SAD COMPARISONS



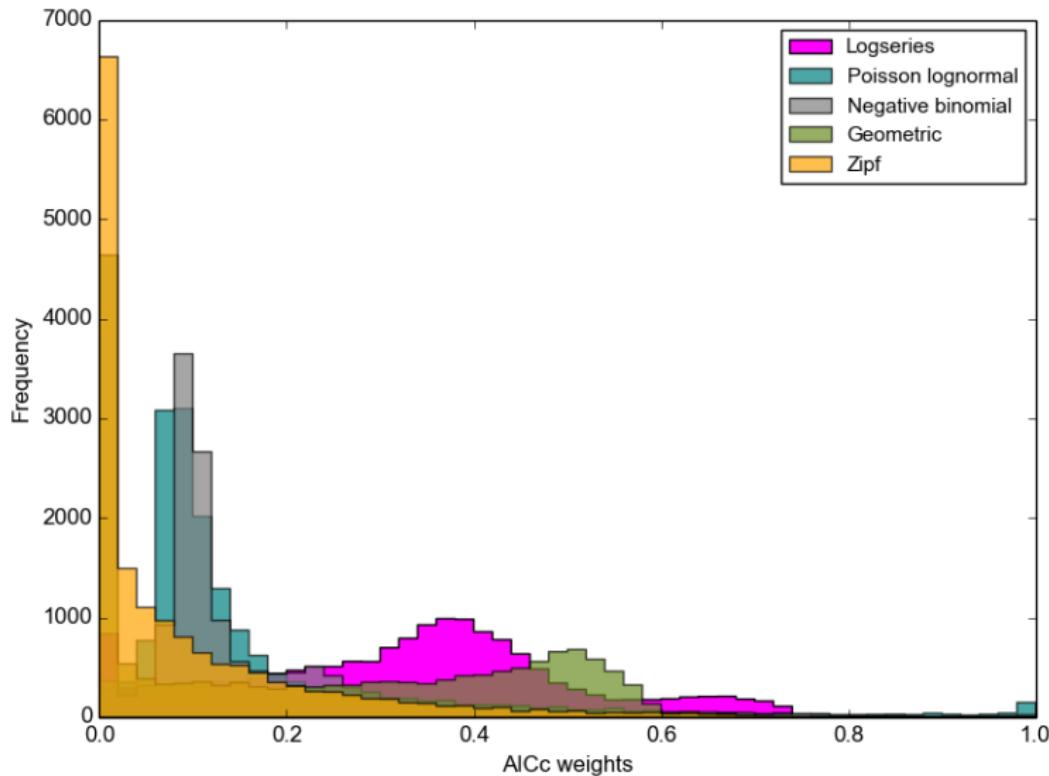
# SAD COMPARISONS



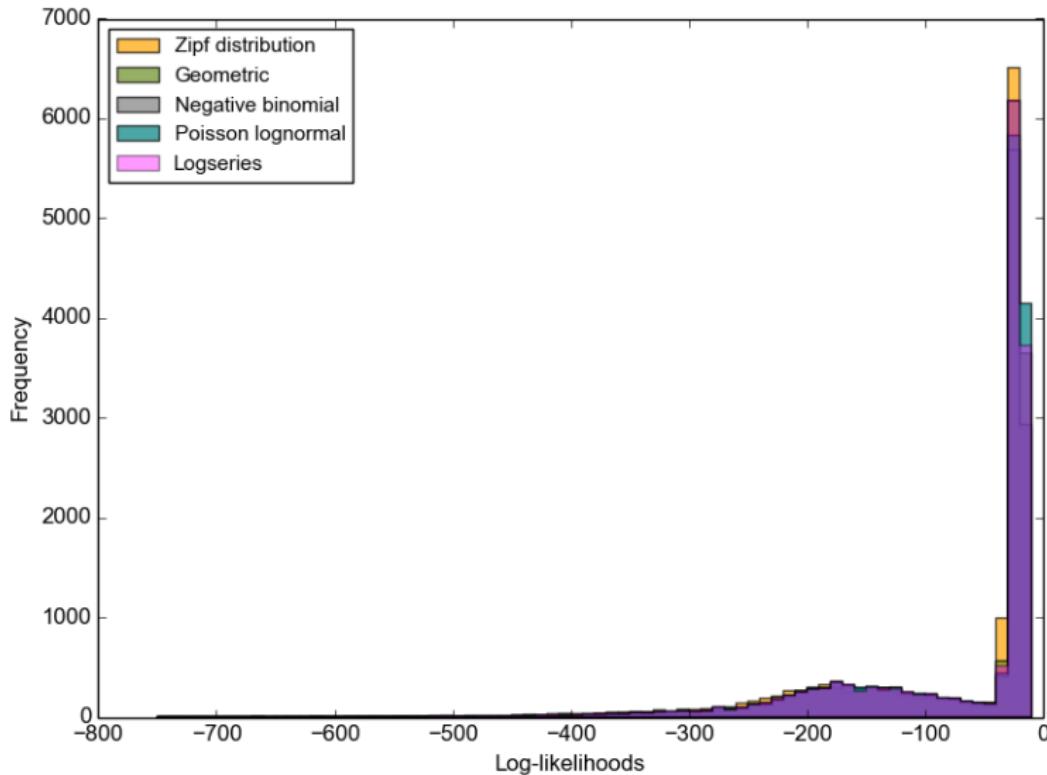
# SAD COMPARISONS



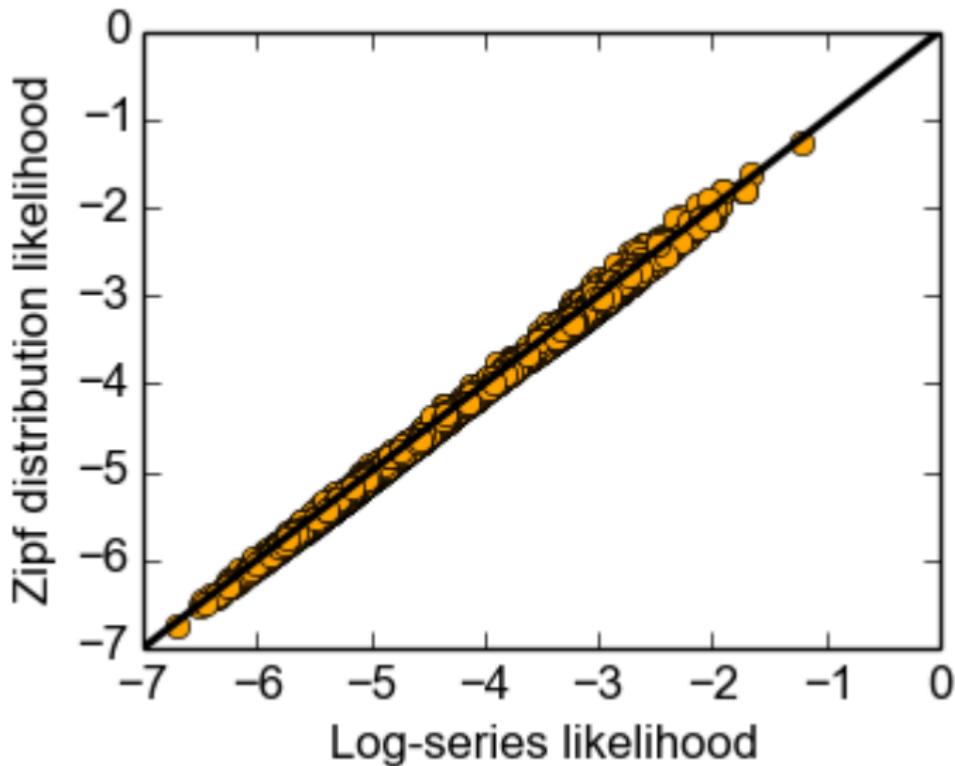
# SAD COMPARISONS



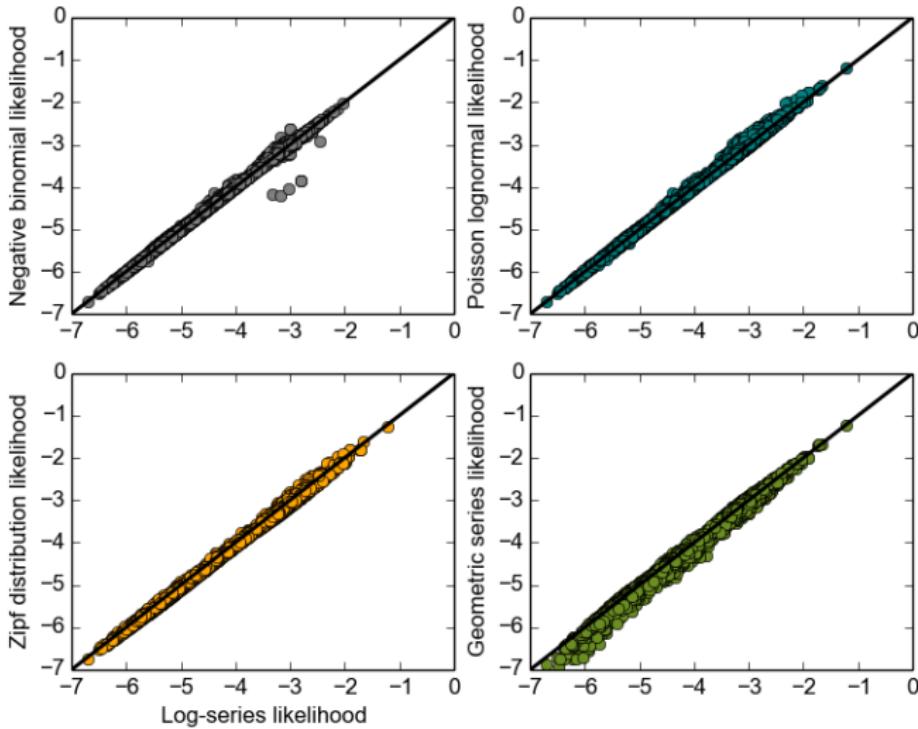
# SAD COMPARISONS



## SAD COMPARISONS



# SAD COMPARISONS



## SAD COMPARISONS

Existing models provide equivalently good absolute fits to empirical data.

- Models with fewer parameters perform better in AIC-based model selection.
- Logseries provides a good naive model for fitting SADs.
  - Produces equivalent likelihoods.
  - Has a single fitted parameter.
  - Easy to fit to empirical data.
  - Best overall model.

## SAD COMPARISONS

Identifying pattern generating mechanisms:

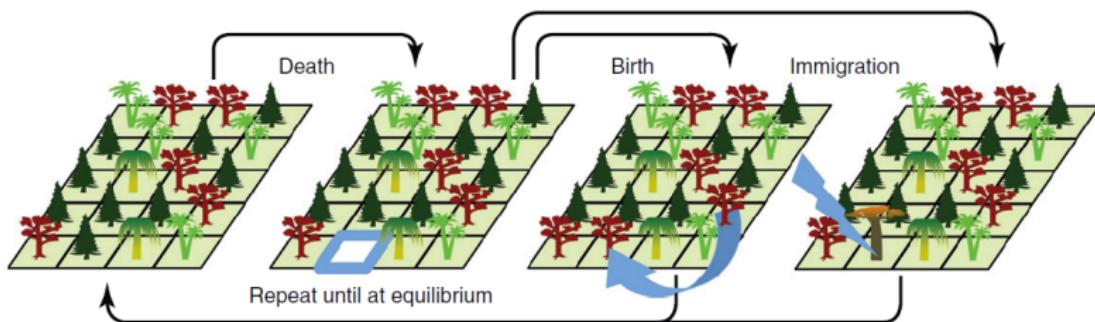
- Compare predictions of different models using multiple macroecological patterns simultaneously.
- Examine scale dependence of pattern.

However, identification of mechanism may not be necessary for prediction.

# NEUTRAL THEORY

Many formulations, but:

- Species & individuals ecologically & demographically equivalent.



Rosindell et al. 2011.

## NEUTRAL ANALYSIS

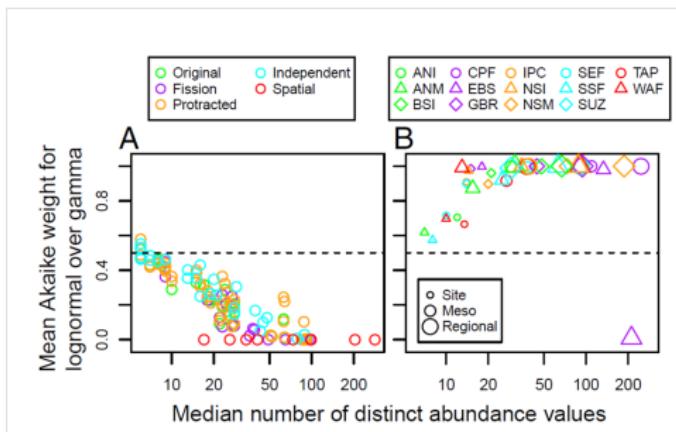
Early tests of neutral theory compared the fit of empirical species abundance distributions to the neutral prediction.

Later tests suggested species abundance comparisons were insufficient for a rigorous test of neutrality.

*However...*

# NEUTRAL ANALYSIS

Connolly et al. 2014 identified non-neutral species abundance distributions in marine communities.



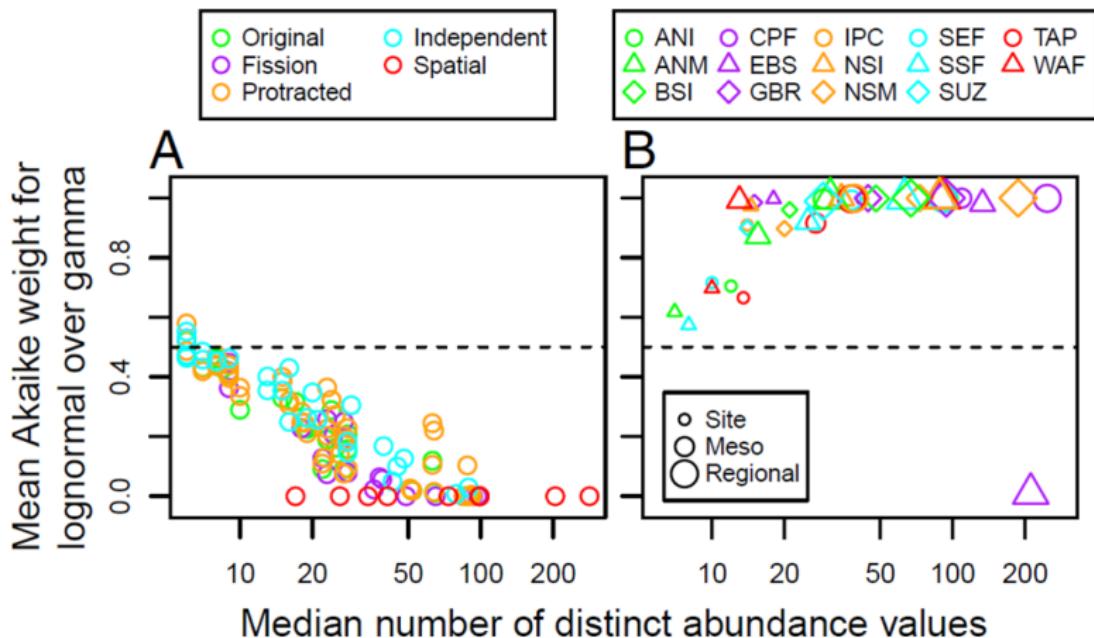
May be a robust method for identifying communities that exhibit non-neutrality.

## NEUTRAL ANALYSIS

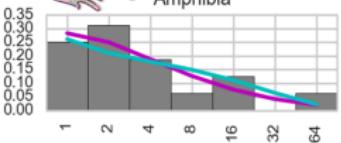
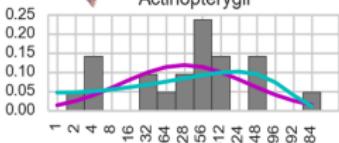
Used the same data and model fitting approach.

Compared a non-neutral model (Poisson lognormal) to a neutral model (negative binomial).

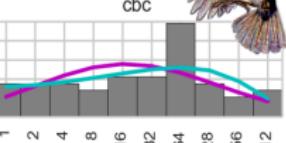
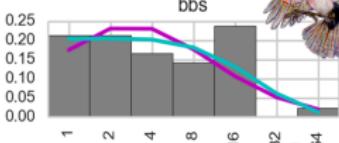
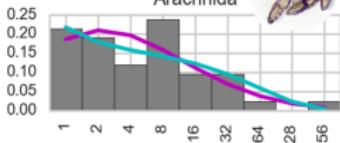
# NEUTRAL ANALYSIS



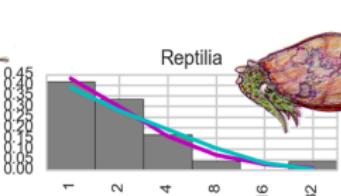
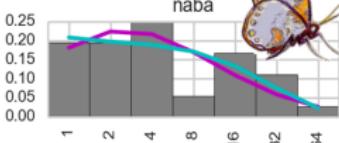
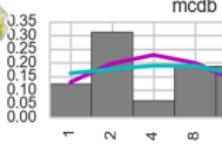
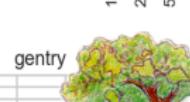
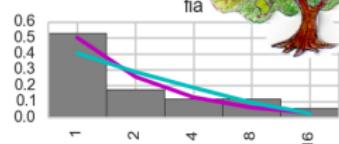
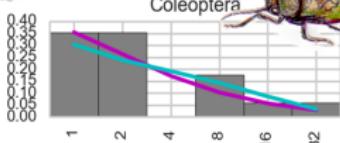
# NEUTRAL ANALYSIS



Arachnida

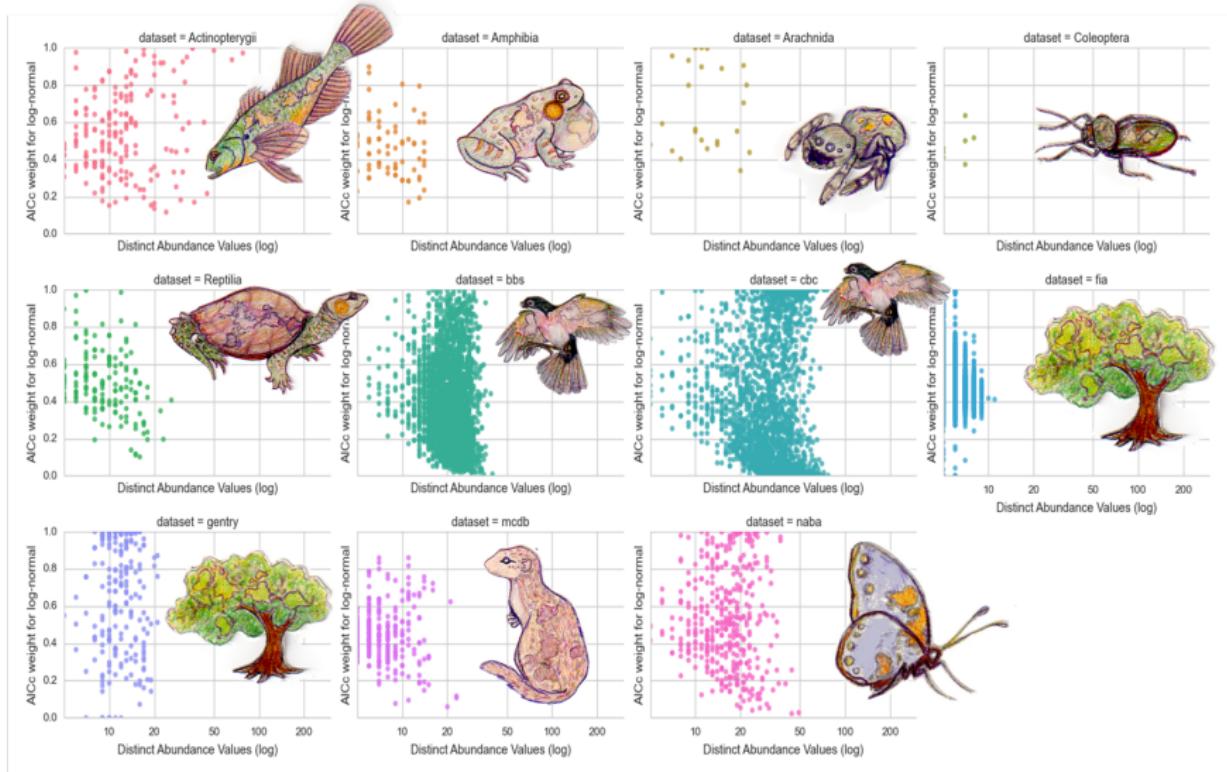


Coleoptera

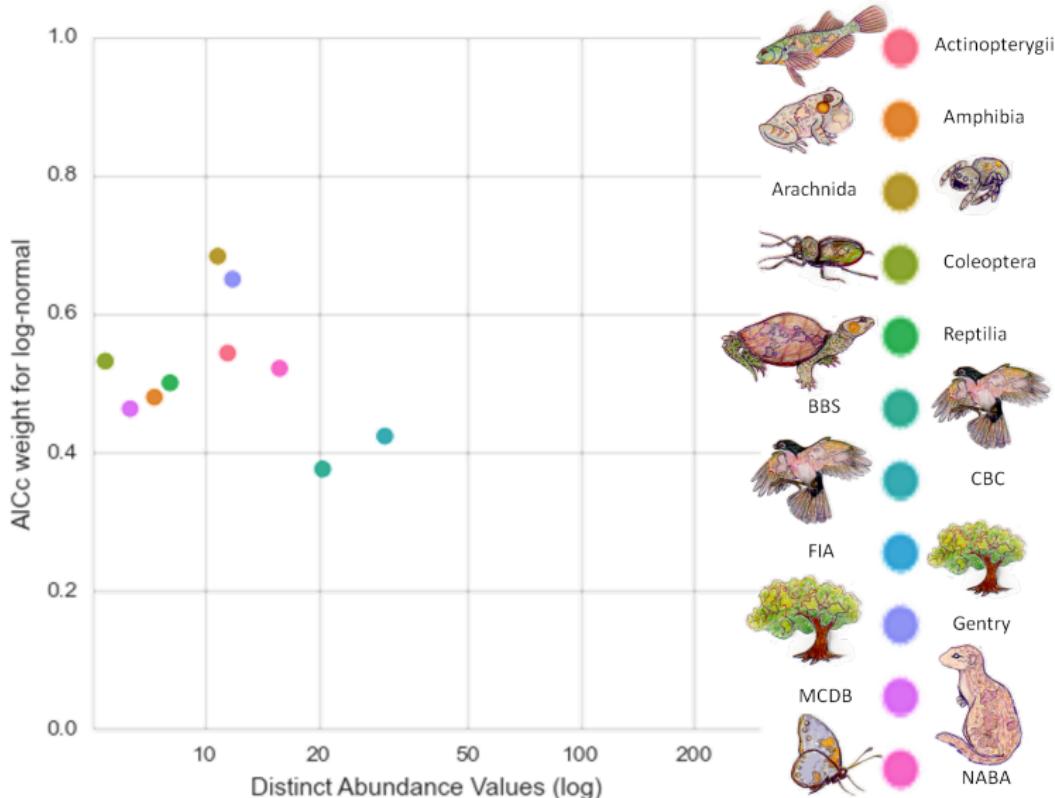


Poisson lognormal  
Negative binomial

# NEUTRAL ANALYSIS



# NEUTRAL ANALYSIS



## NEUTRAL ANALYSIS

Difficult to identify a clear winning model.

- Results consistent with our species abundance distribution model comparisons.
- Results different from Connolly et al. 2014.
  - Non-neutral model outperforms the neutral model in marine systems.
  - Our results suggest marine systems more generally approximated by non-neutral dynamics; terrestrial systems more variable between neutral and non-neutral dynamics.

# CONCLUSIONS

Challenging to infer process from species abundance distributions alone.

- Multiple mechanisms proposed for each SAD formulation.
- Broad model categorization (i.e. neutral or non-neutral) may be more productive.
- May not be one single suite of processes that dominates.

# CONCLUSIONS

Challenges in identifying mechanism among datasets.

- Biological vs. non-biological differences (spatial structuring, sampling intensity).
- Diverse data removes uncertainty about non-biological pattern generating mechanisms.
- Even with a great deal of data, identifying mechanism is still challenging.

# CONCLUSIONS

## Predictive macroecology

- Traditional approach is pattern to process to prediction.
- May be possible to generate robust ecological predictions from general patterns.
- Process and prediction may be two separate research goals.

## ACKNOWLEDGEMENTS

### Funding sources:

- USU Department of Biology
- Intellectual Ventures, private funding to Morgan Ernest
- National Science Foundation CAREER Grant to Ethan White
- Gordon & Betty Moore Foundation's Data-Driven Discovery Initiative Grant to Ethan White.
- USU Graduate School Dissertation Fellowship

# ACKNOWLEDGEMENTS

## Weecologists past, present, & future



(especially Xiao Xiao & Ken Locey (creator of the whiteboard))

## ACKNOWLEDGEMENTS

Dr. Thomas Price & USU Student Health Center.

Tea, heating pads, & the Flint Hills of Kansas.

A very supportive husband & family.

Publicly available data, & the citizen scientists that make that possible.

# ACCESSIBILITY

This dissertation brought to you by:

## Disability accommodations

- Remote access & participation.
- Computational tools & tricks.
  - Version control (GitHub).
  - Publicly available data.
  - Programming skills (data manipulation & analysis).

# QUESTIONS?

