

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

Feel free to:



Copy, share, adapt, or re-mix;



Photograph, film, or broadcast;



Blog, live-blog, or post video of;

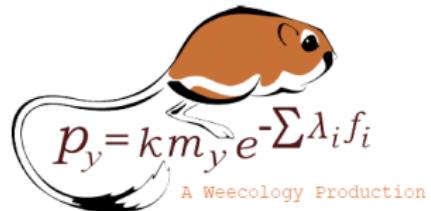
Provided that:



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

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

Elita Baldridge  
@elitabaldridge



# OPEN SCIENCE

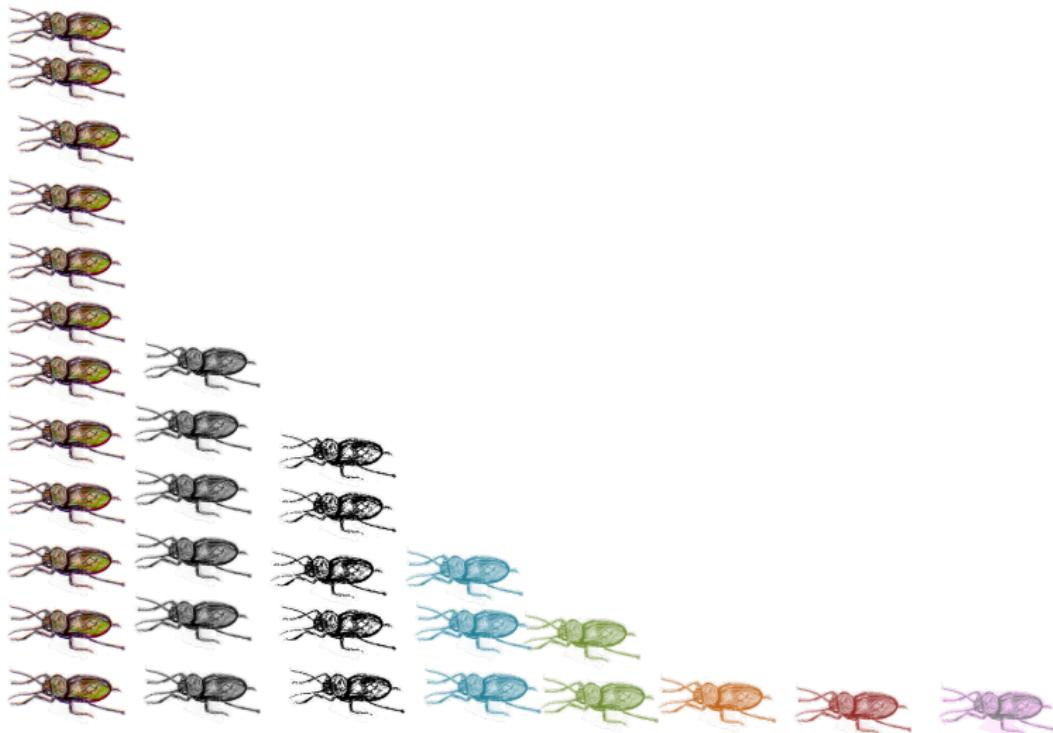
- Code:
  - [github.com/embaldridge](https://github.com/embaldridge)
  - [github.com/weecology](https://github.com/weecology)
- Data: [figshare.com](https://figshare.com)



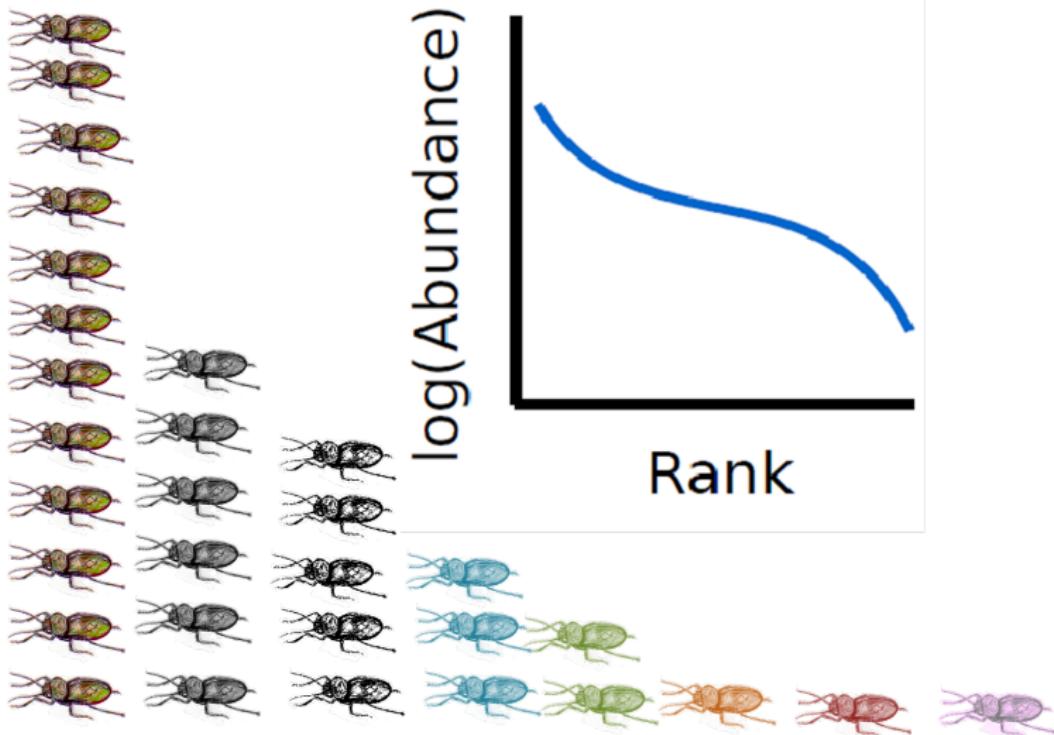
**figshare**  
credit for all your research



# LET'S BEGIN WITH AN EXAMPLE...



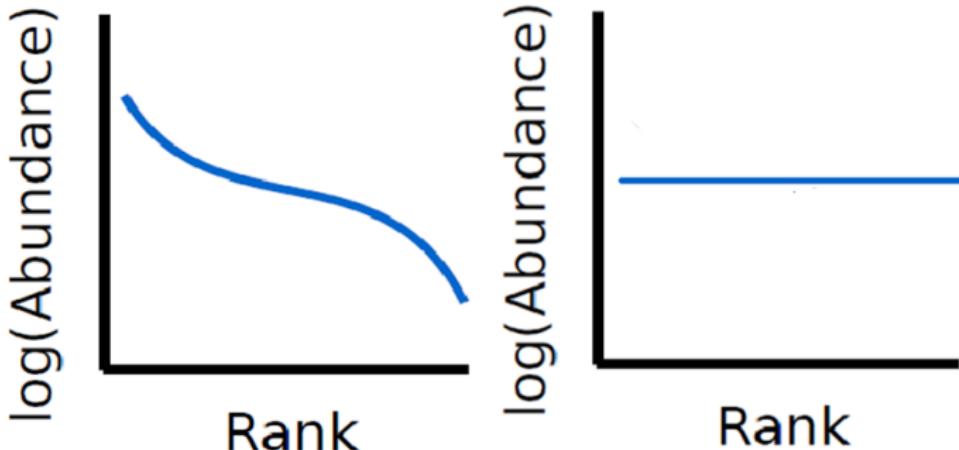
## LET'S BEGIN WITH AN EXAMPLE...



# PATTERN & PROCESS.

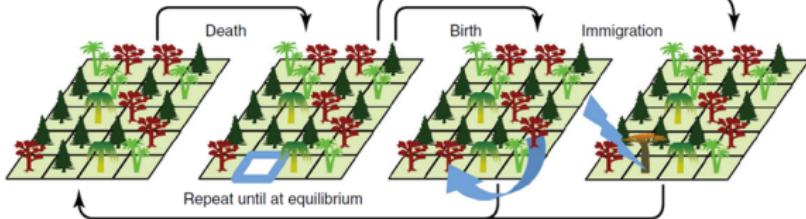
## Species abundance distribution (SAD)

- Reveals pattern-generating mechanisms of community structure.



# PATTERN & PROCESS.

## PROCESS



$\log(\text{Abundance})$

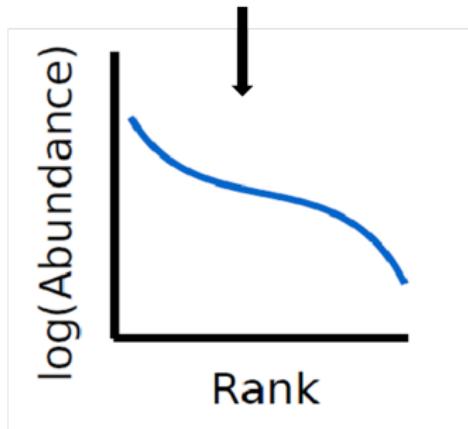


Rank  
**PATTERN**

# PATTERNS & PROCESS.

## Process

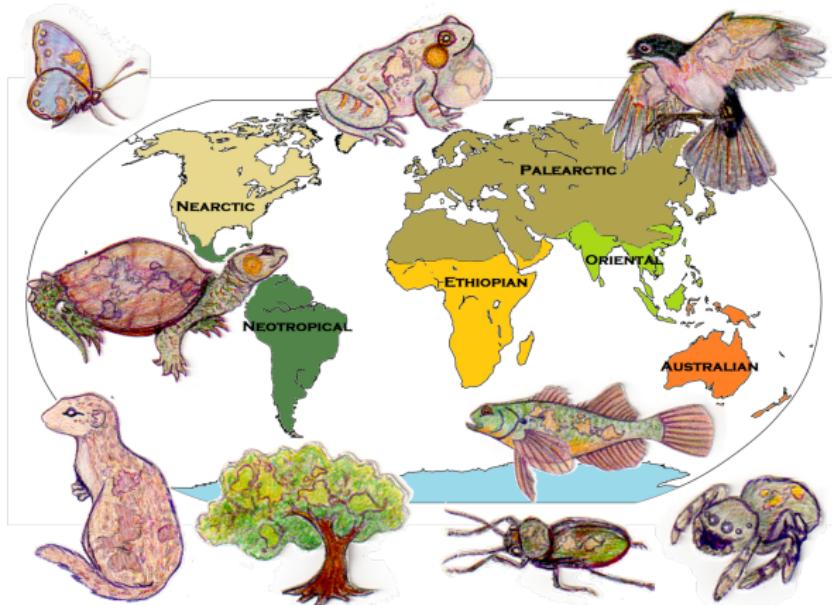
**COMPETITION    NICHES    DISPERSAL**  
**NUTRIENTS    NEUTRALITY    ETC.**



## Pattern

# MACROECOLOGY

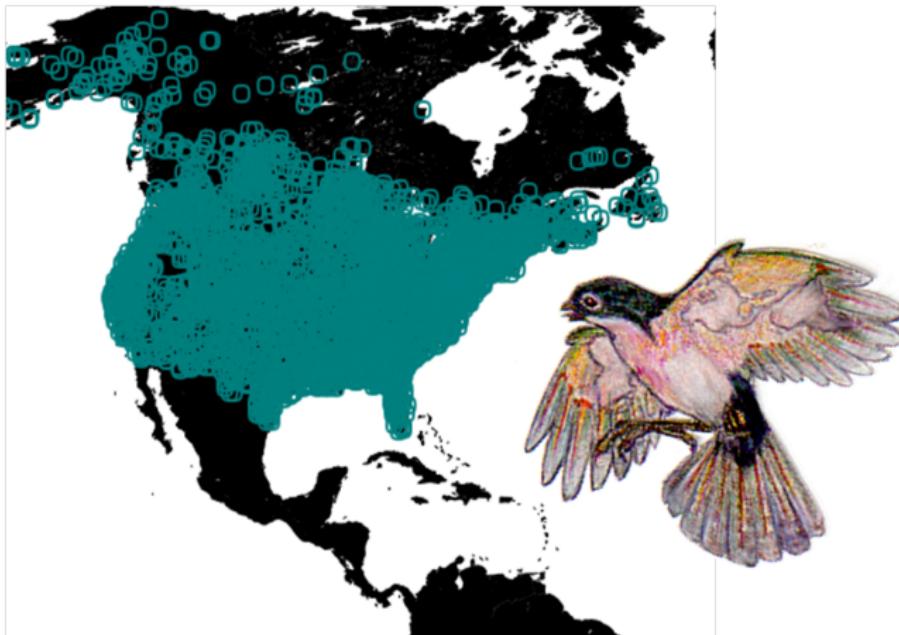
One approach for identifying general ecological patterns & processes.



# TRADITIONAL APPROACH



## BROADER APPROACH



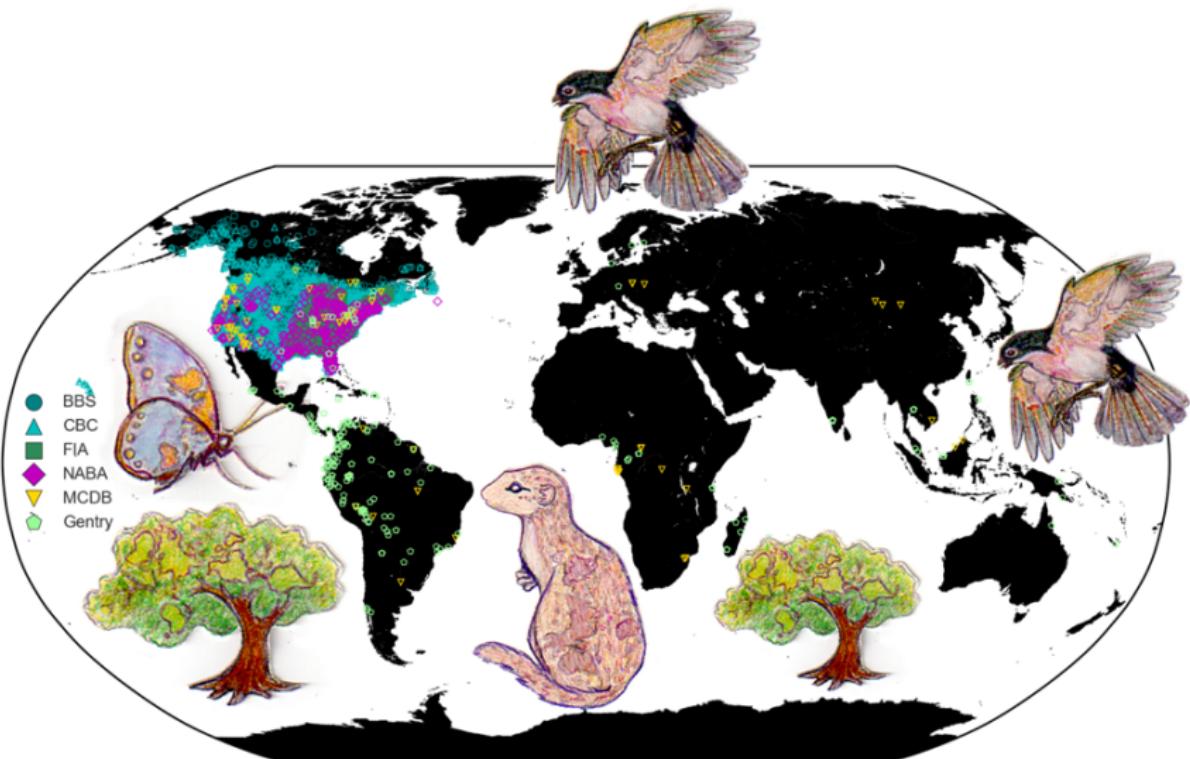
# PATTERN & PROCESS; SIGNAL & NOISE



# DATA-INTENSIVE ECOLOGY

- Leveraging existing ecological data.
- The species abundance distribution.
- The statistical approach.
- The mechanistic approach.

# CURRENT DATA



# MACROECOLOGICAL DATA

## Challenges of macroecology

- Lack of identification of process.

## Best practice recommendations

- Test with multiple taxonomic groups/ecosystems.

Plenty of data in the literature.

# THE RULES OF ECOINFORMATICS

Garbage in, garbage out.

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

# BUILDING A DATABASE

Record decisions at all steps:

*Metadata are important.*

## ABUNDANCE DATABASE

Inclusion criteria:

- Quantitative abundances (counts).
- Complete sampling.
- Raw data (not heavily processed).
- Observational.

# ABUNDANCE DATABASE

## Variables collected

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

TABLE : List of variables collected.

# DATA WRANGLING

- Fire and a Tallgrass Prairie Reptile Community: Effects on Relative Abundance and Seasonal Activity

^ v p. 14 (4 of 10) < >

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

# DATA WRANGLING

Three LibreOffice Calc windows are displayed side-by-side, illustrating data wrangling steps:

- Species\_abundances.csv - LibreOffice Calc**: A data frame showing species abundance across sites. Row 20 has a highlighted cell in column E.
- Sites\_table\_abundances.csv - LibreOffice Calc**: A table mapping site IDs to names and locations. Site 10 is highlighted.
- Citations\_table\_abundance.csv - LibreOffice Calc**: A bibliography table with columns for citation ID, authors, year, and title. Citation 1 is highlighted.

**Species\_abundances.csv Data (approximate values):**

	A	B	C	D	E	F	G	H
1	Class	Family	Genus	Species	Relative_abundance	Abundance	Site_ID	Citation
2	Reptilia	Pitophis	catenifer		0	0	1	1
3	Reptilia	Lampropeltis	calligaster		0	0	1	1
4	Reptilia	Storena	dekayi		0	0	1	1
5	Reptilia	Eumeces	septentrionalis		0	0	1	1
6	Reptilia	Eumeces	obsoleteus	0.28	1	1	1	
7	Reptilia	Elaphe	emoryi	1.72	2	1	1	
8	Reptilia	Tropidoclonion	lineatum	0.57	2	1	1	
9	Reptilia	Lampropeltis	triangulum	0.85	3	1	1	
10	Reptilia	Elaphe	obsoleta	1.14	4	1	1	
11	Reptilia	Thamnophis	sirtalis	1.42	5	1	1	
12	Reptilia	Lampropeltis	getula	0.66	5	1	1	
13	Reptilia	Ophisaurus	attenuatus	4.83	17	1	1	
14	Reptilia	Coluber	constrictor	9.37	33	1	1	
15	Reptilia	Pitophis	catenifer	0	0	2	1	
16	Reptilia	Tropidoclonion	lineatum	0	0	2	1	
17	Reptilia	Lampropeltis	calligaster	0	0	2	1	
18	Reptilia	Storena	dekayi	0	0	2	1	
19	Reptilia	Lampropeltis	triangulum	0.13	1	2	1	
20	Reptilia	Elaphe	obsoleta	0.13	1	2	1	
21	Reptilia	Eumeces	septentrionalis	0.13	1	2	1	
22	Reptilia	Eumeces	obsoleteus	0.26	2	2	1	
23	Reptilia	Lampropeltis	getula	0.77	5	2	1	
24	Reptilia	Thamnophis	sirtalis	1.59	12	2	1	
25	Reptilia	Elaphe	emoryi	0.38	13	2	1	
26	Reptilia	Ophisaurus	attenuatus	2.91	22	2	1	
27	Reptilia	Coluber	constrictor	5.82	44	2	1	
28	Reptilia	Tropidoclonion	lineatum	0	0	3	1	
29	Reptilia	Storena	dekayi	0	0	3	1	
30	Reptilia	Eumeces	obsoleteus	0	0	3	1	

**Sites\_table\_abundances.csv Data (approximate values):**

A	B	C	D	E	F	G	H
3	1996	1	Konza A	Nearctic	20	hectares, burned 1980	1985
5	1996	1	Konza B	Nearctic	80	hectares, burned in 1980.	1986
7	1996	1	Konza D	Nearctic	36	hectares, annual burn 1972-197	1973
9	1996	1	Konza C	Nearctic	90	hectares, burned 1980, 1985,	1
2	1995	1	Konza A	Nearctic	80	hectares, burned in 1980.	1986
4	1995	1	Konza B	Nearctic	36	hectares, annual burn 1972-197	1973
6	1995	1	Konza C	Nearctic	90	hectares, annual burn 1980, 1985,	1
8	1995	1	Konza D	Nearctic	36	hectares, burned 1980, 1991	
1	1994	1	Konza A	Nearctic	90	hectares, burned 1980, 1985,	1
10	1980	2	Treatment A	Neotropical	Artificial fall	removals monthly for	

**Citations\_table\_abundance.csv Data (approximate values):**

A	B	C	D
Citation_ID	Authors	Yr	Title
1	Cavitt, John F.	2000	Fire and a tallgrass prairie reptile community: Effects on re
2	Bullman, T.L. and G.	1982	Abundance and community structure of forest floor spiders
3	Schlosser, I.J.	1985	Flow regime, juvenile abundance, and the assemblage stru
5	Jones, K. B.	1981	Effects of grazing on lizard abundance and diversity in We
6	Grossman, G.D.	1982	Dynamics and organization of a rocky intertidal fish assem
4	Brandt, A.	1997	Abundance, diversity and community patterns of epibenthic
8	Ortizchillo, W and En	1982	Responses in abundance and diversity of cornfield carabid
10	Petterson, R.B.	1996	Effects of forestry on the abundance and diversity of arbo
12	Menke, S.B.	2003	Lizard community structure across a grassland- creosote b

# DATA WRANGLING

The image shows four terminal windows side-by-side, each displaying a different CSV file or dataset.

- Metadata.txt:** A plain text file containing an introduction and metadata class descriptions.
- Species\_abundances.csv:** A CSV file with columns: Class, Family, Genus, Species, Relative\_abundance, Abundance, Site\_ID, Citation. The data includes various reptile species and their abundance across different sites.
- Sites\_table\_abundances.csv:** A CSV file with columns: Site\_ID, Collection\_Year, End\_Collection, Citation\_ID, Site\_Name, Biogeographic\_region.
- Citations\_table\_abundances.csv:** A CSV file with columns: Citation\_ID, Authors, Yr, Title, Journal, Issue, Pages. It lists scientific publications related to the study.

**INTRODUCTION**  
This dataset was developed to provide a source of abundance data for groups that do not have extensive compilations of abundance data.

There are several caveats to the use of this database. Abundance has been recorded as the raw abundance or the relative abundance, depending on what was available from the original source. Abundance is the total number of individuals captured, relative abundance is the total number of individuals captured for a single species/ total number of individuals of all species.

METADATA CLASS I. DATA SET DESCRIPTIONS

A. Data set identity:

- Title: MiscAbundance

B. Data set identification code:

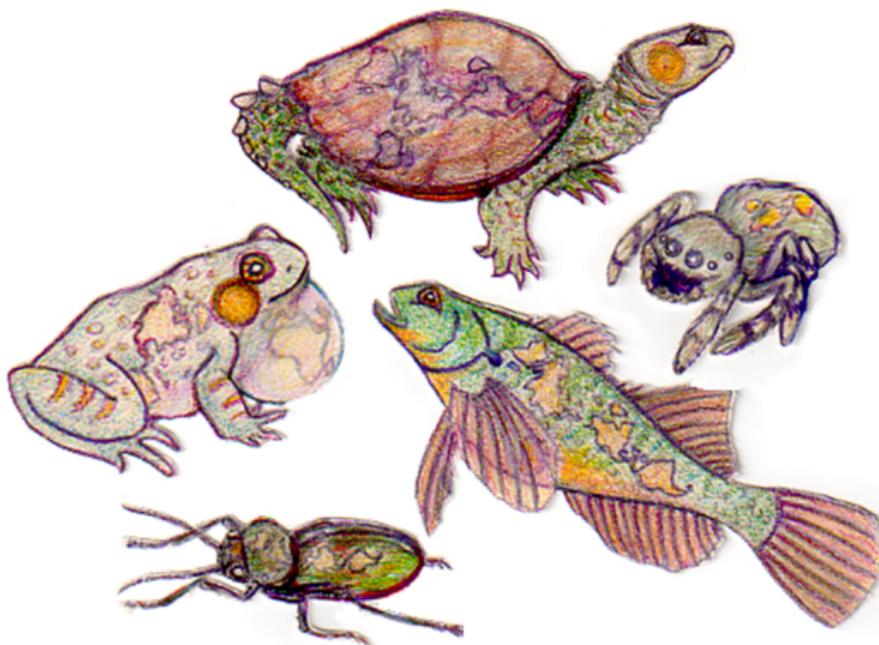
1. Abundance data: Species\_abundances.csv
2. Sites data file : Sites\_table\_abundances.csv
3. Reference file: Citations\_table\_abundances.csv

C. Data set description

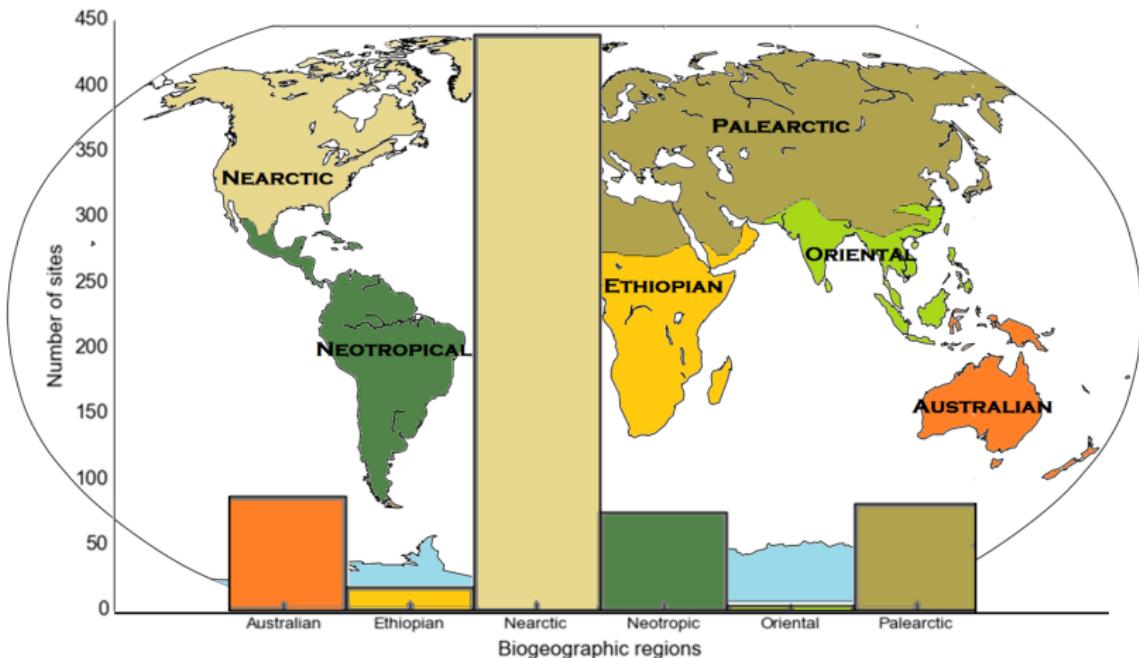
Principal Investigators:

Plain Text • Tab Width: 8 • Ln 6, Col 44

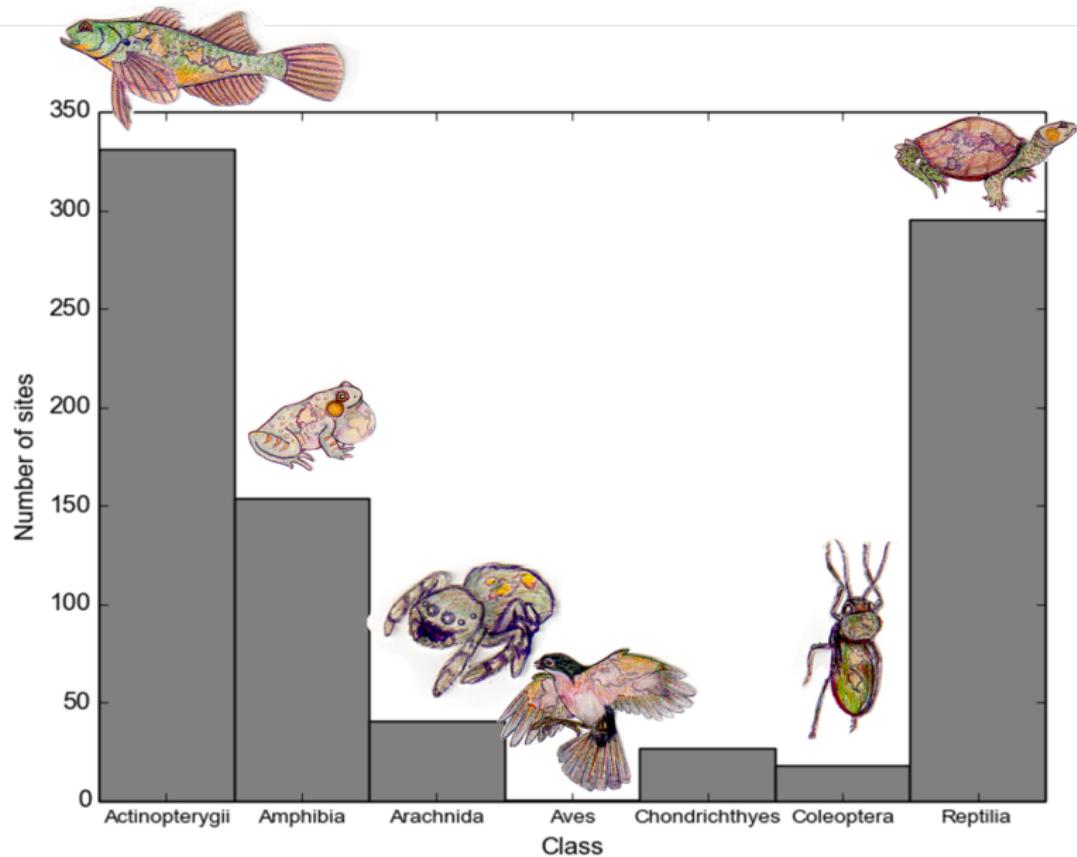
# ABUNDANCE DATABASE



# ABUNDANCE DATABASE



# ABUNDANCE DATABASE



## DATA AVAILABILITY

Public & open access through figshare.  
EcoData Retriever importable.

(<http://figshare.com>)

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

sad\_data =

ecoretriever::fetch('MiscAbundanceDB')



# DATA



# DATA



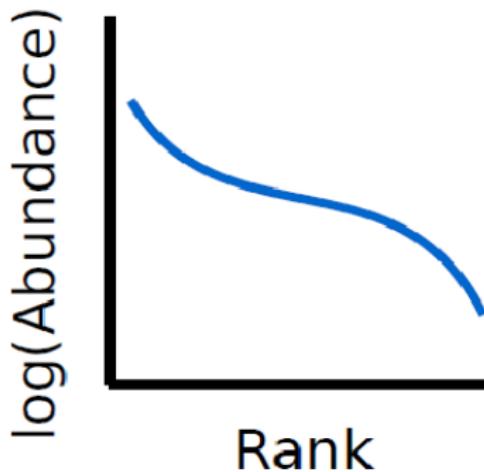
# SAD COMPARISONS

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.

## COMMONNESS & RARITY

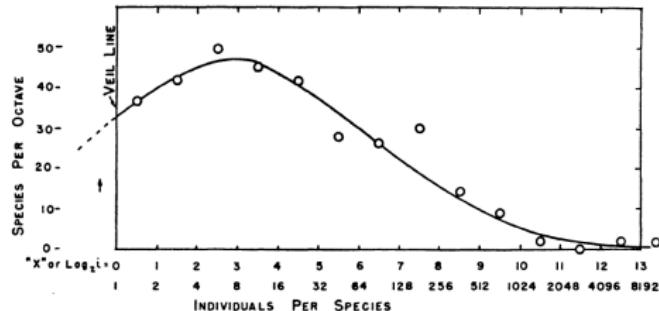
The species abundance distribution:



Many models of the species abundance distribution (SAD).

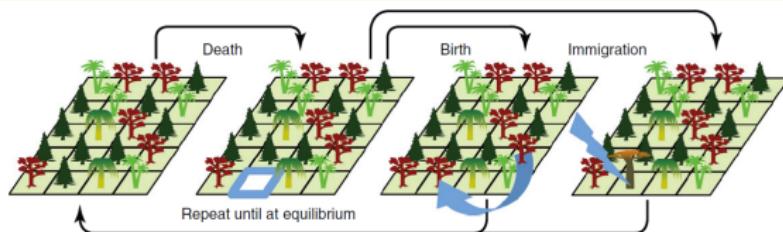
# SAD MODELS

## Statistical description



Preston 1962a.

## Process-based



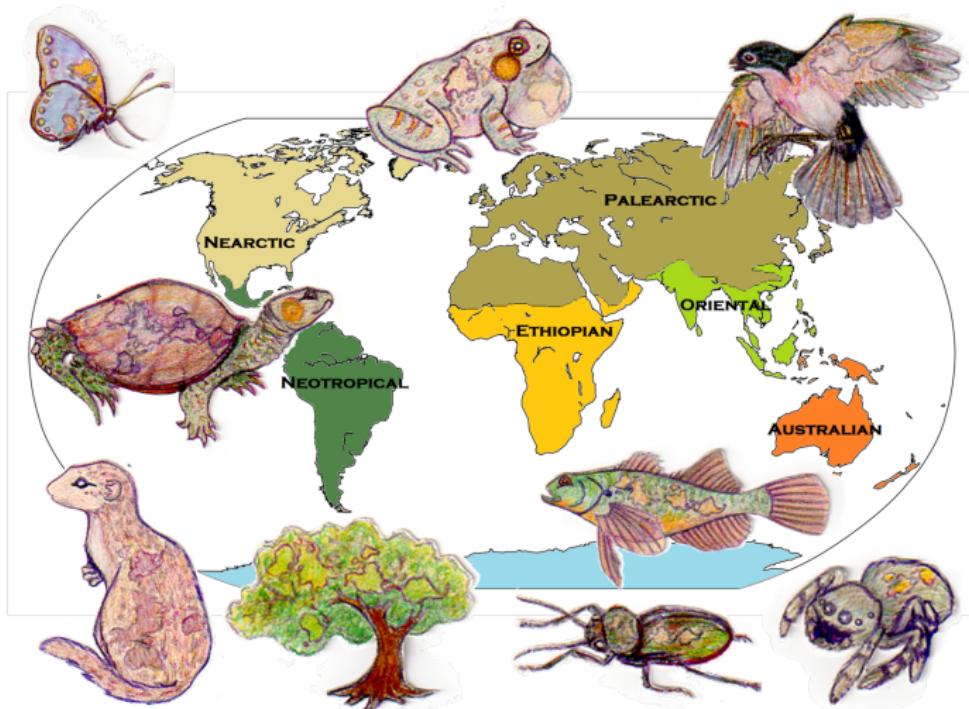
Rosindell et al. 2011.

## SAD COMPARISONS: THE STATISTICAL APPROACH.

Most comparisons of the different models:

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

# SAD COMPARISONS



# 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.
  - For a given model, estimates model parameters that provide the most likely characterization of the data.
  - Best practice for fitting species abundance distributions. Matthews & Whittaker 2014.

# SAD COMPARISONS

## Analysis:

- Likelihood based model selection to compare the fits of the different models.
  - How well does the model describe the data?

# SAD COMPARISONS

## Analysis:

- Model comparison with corrected Aikaike Information Criterion (AICc) weights.
  - How well does the model describe the data relative to the number of parameters?
- The best fitting model had the greatest AICc weight.

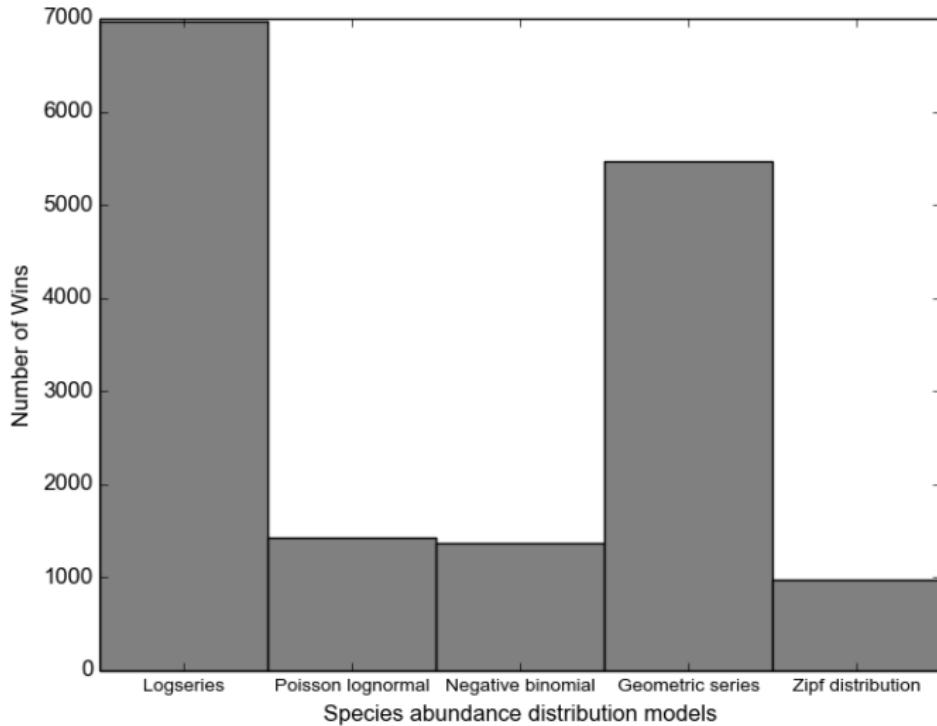
# SAD COMPARISONS

Computational tools:

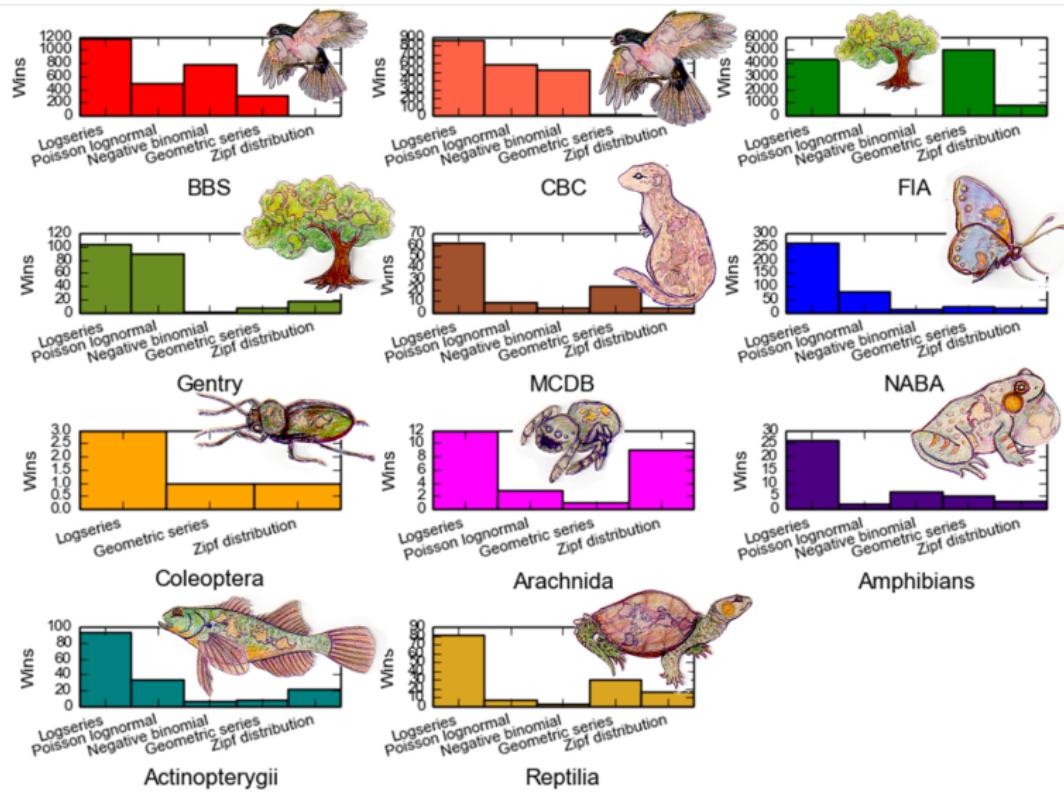
- Model fitting, log-likelihood, & AICc: macroecotools Python package.  
(<https://github.com/weecology/macroeccotools>)
- All code & majority of data are publicly available.  
(<https://github.com/weecology/sad-comparison>)



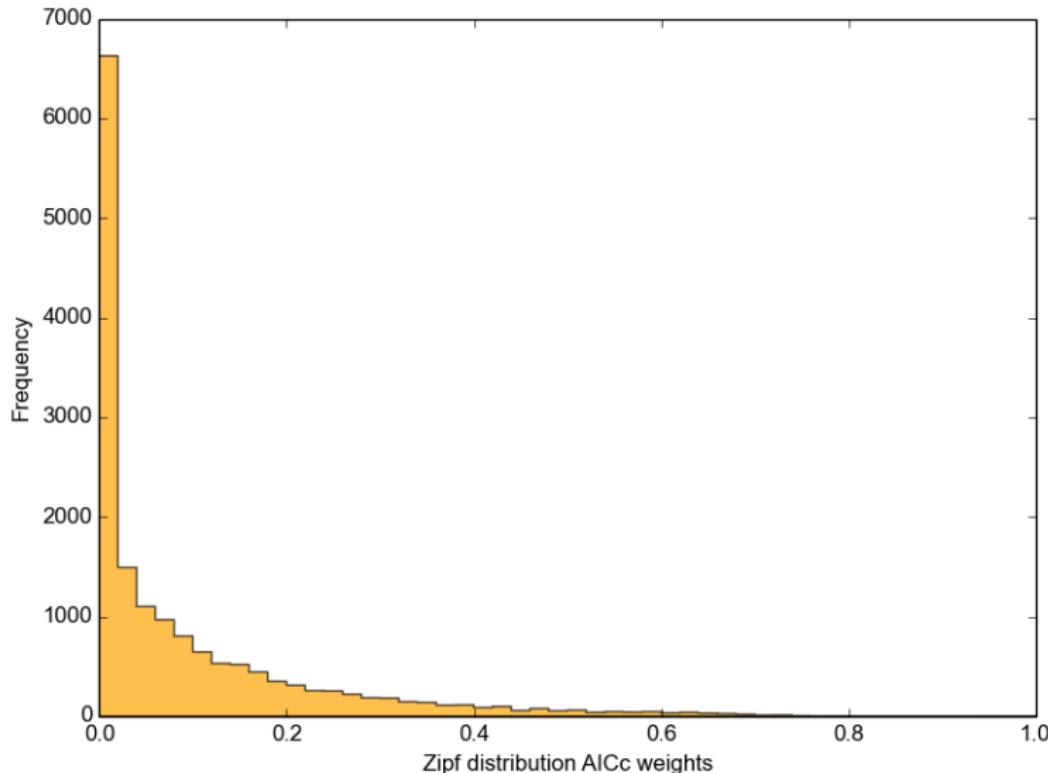
# SAD COMPARISONS



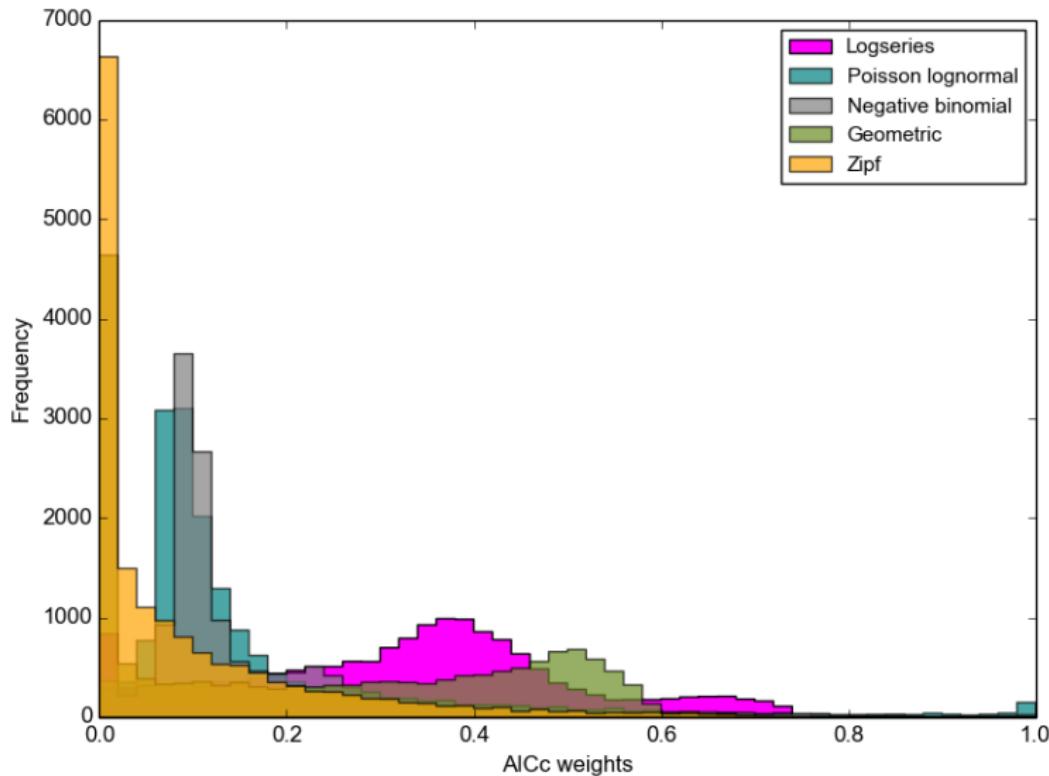
# SAD COMPARISONS



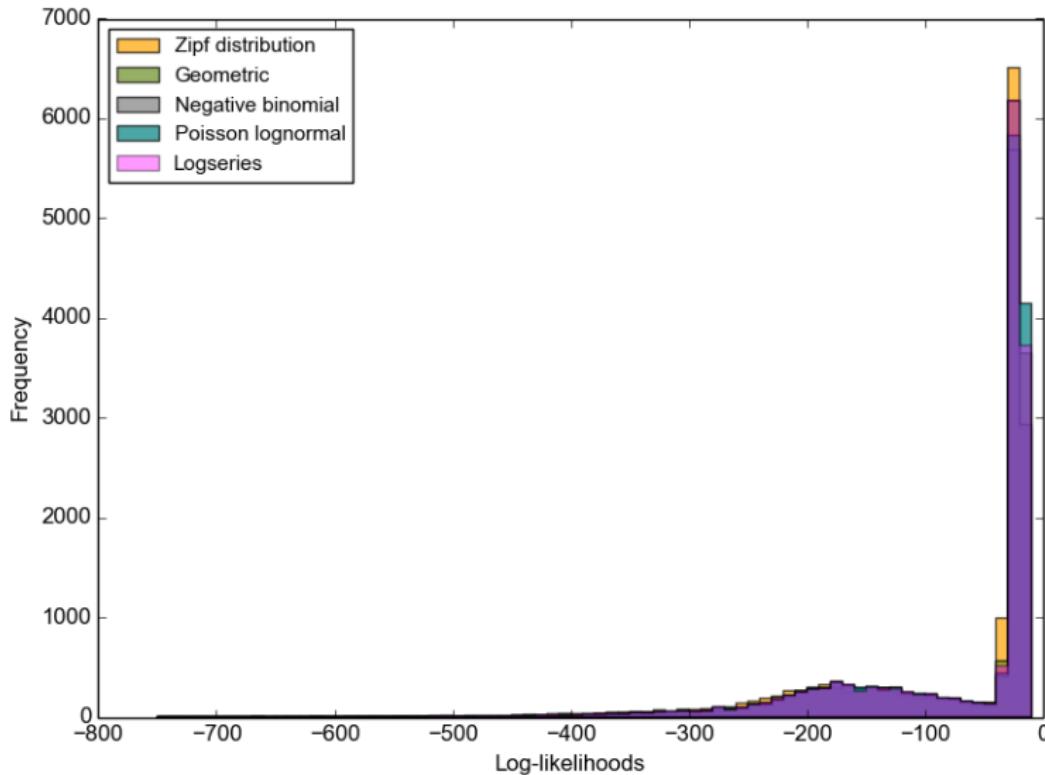
# SAD COMPARISONS



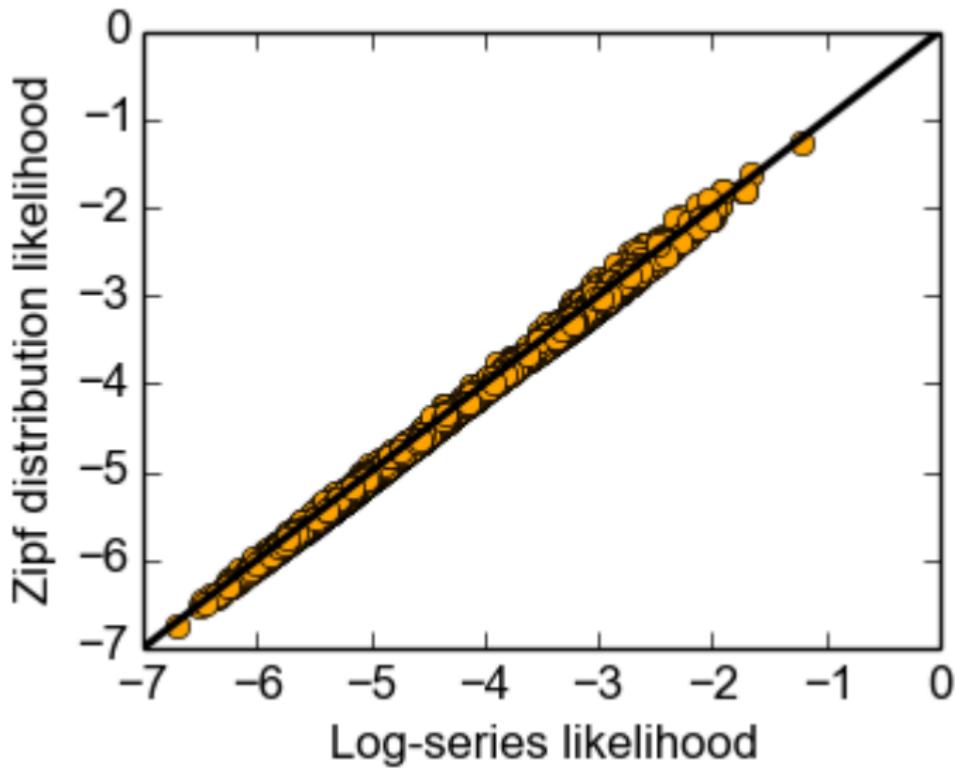
# SAD COMPARISONS



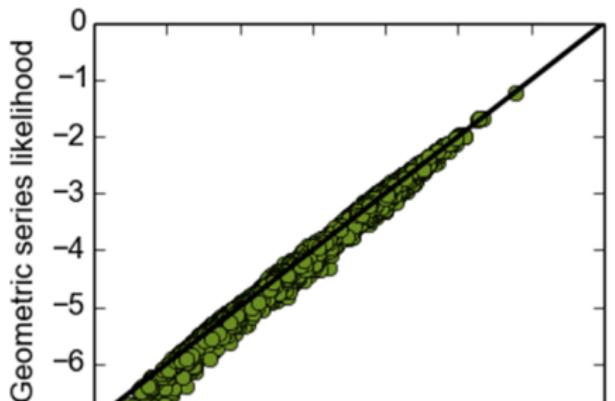
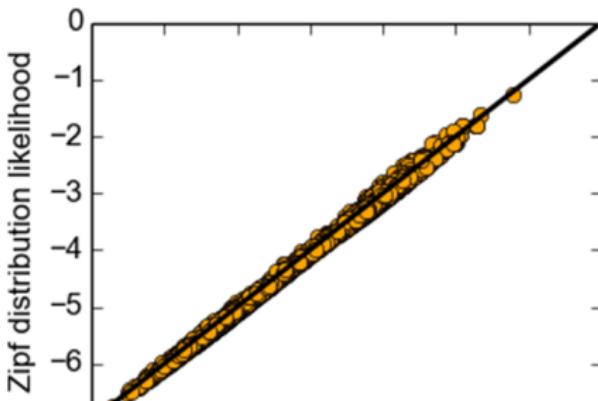
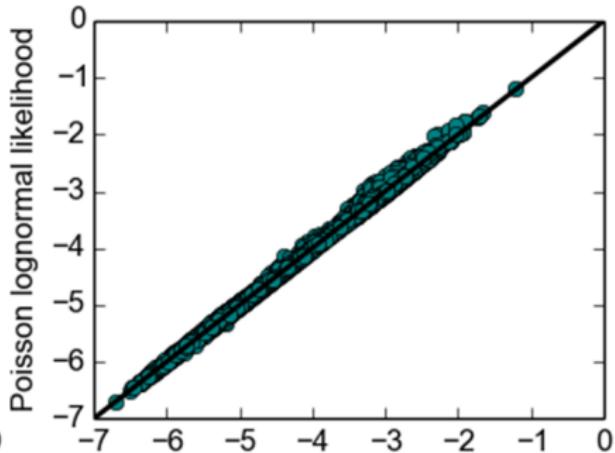
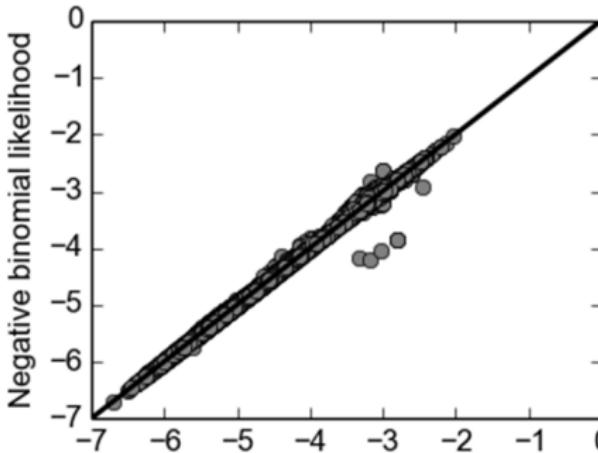
# SAD COMPARISONS



## SAD COMPARISONS



# SAD COMPARISONS

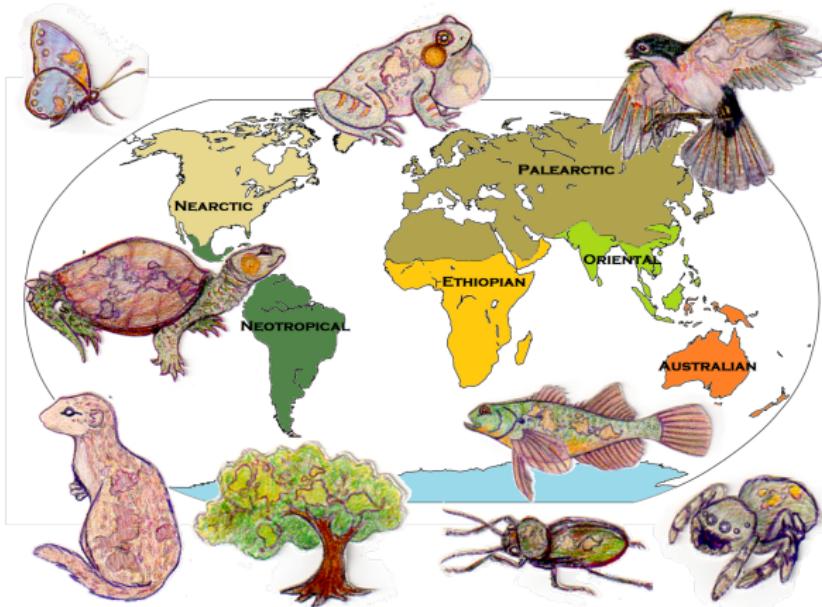


## SAD COMPARISONS

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.

# IDENTIFYING PROCESS:

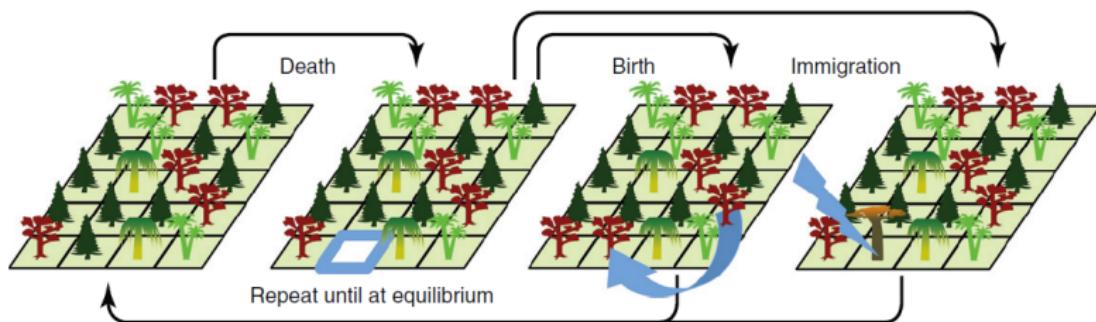


- Examine scale dependence of pattern.
- More general approach to process?

# NEUTRAL PROCESSES

Many formulations of neutral theory, but:

- Species & individuals ecologically & demographically equivalent.



Rosindell et al. 2011.

# NON-NEUTRAL PROCESSES

**COMPETITION    NICHES    DISPERSAL**  
**NUTRIENTS    ETC.**

Many specific mechanisms, but:

- Shape of the abundance distribution due to differences among species.

## THE MECHANISTIC APPROACH: NEUTRAL ANALYSIS.

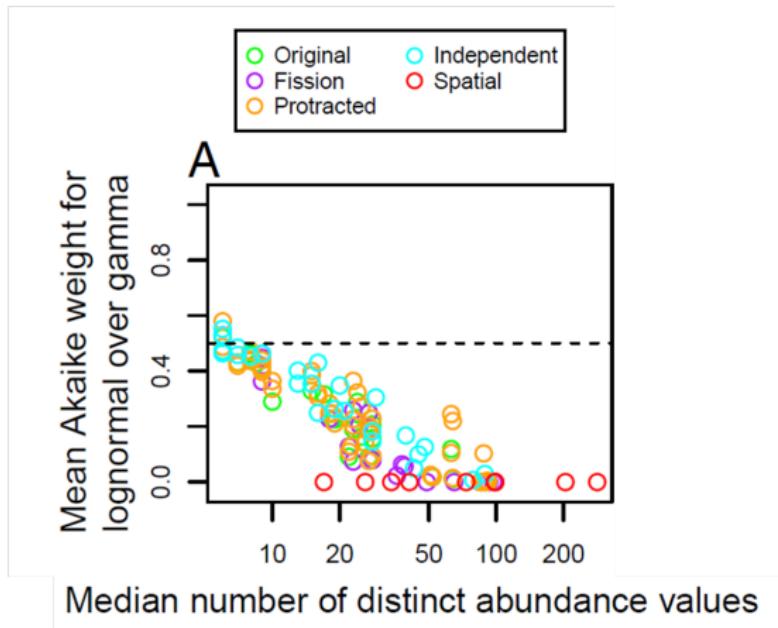
Early tests of neutral theory compared the fit of empirical species abundance distributions to the neutral prediction.

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

*However...*

# NEUTRAL ANALYSIS

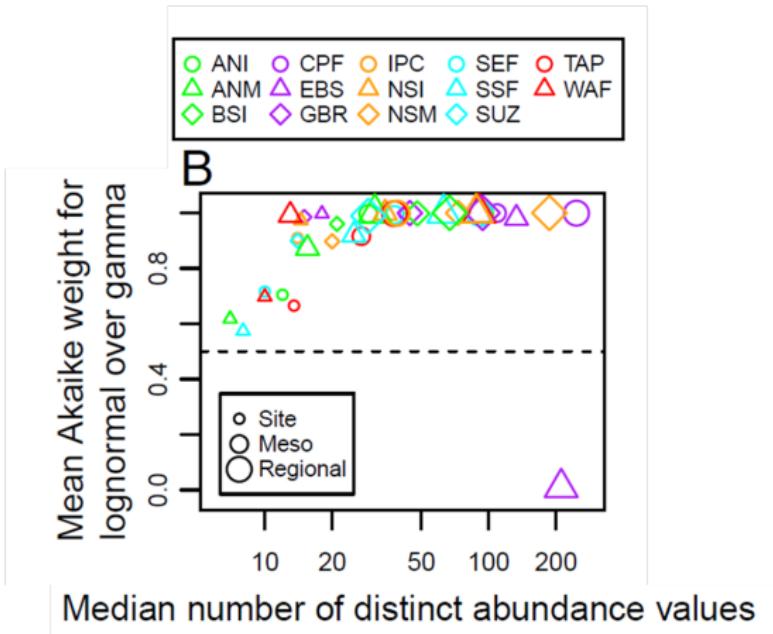
Connolly et al. 2014 simulated neutral communities.



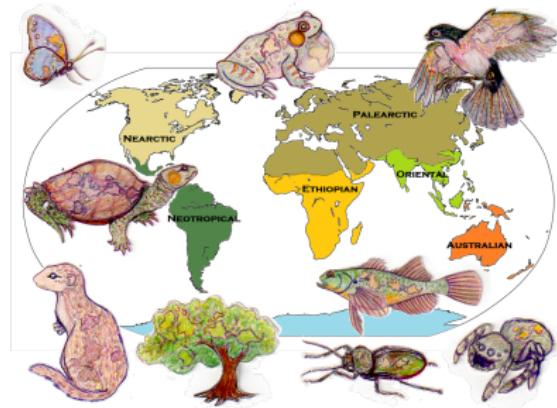
Identified a signal of neutrality.

## NEUTRAL ANALYSIS

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



# SAD COMPARISONS



Used the same data and model fitting approach.

# NEUTRAL ANALYSIS

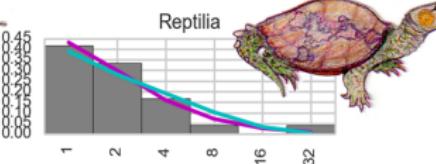
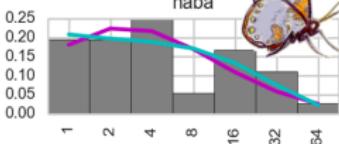
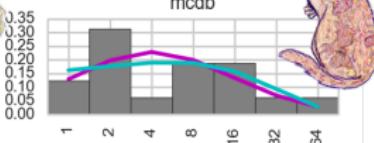
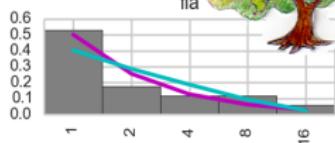
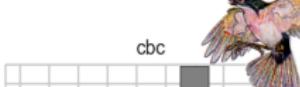
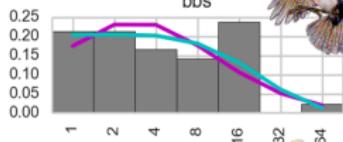
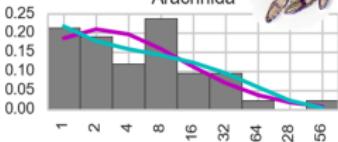
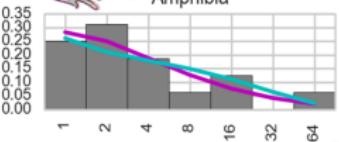
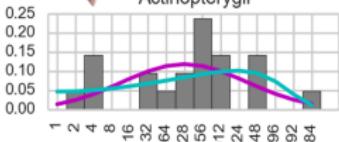
Non-neutral model (Poisson lognormal)

- Describes communities generated by non-neutral processes.

Neutral model (negative binomial).

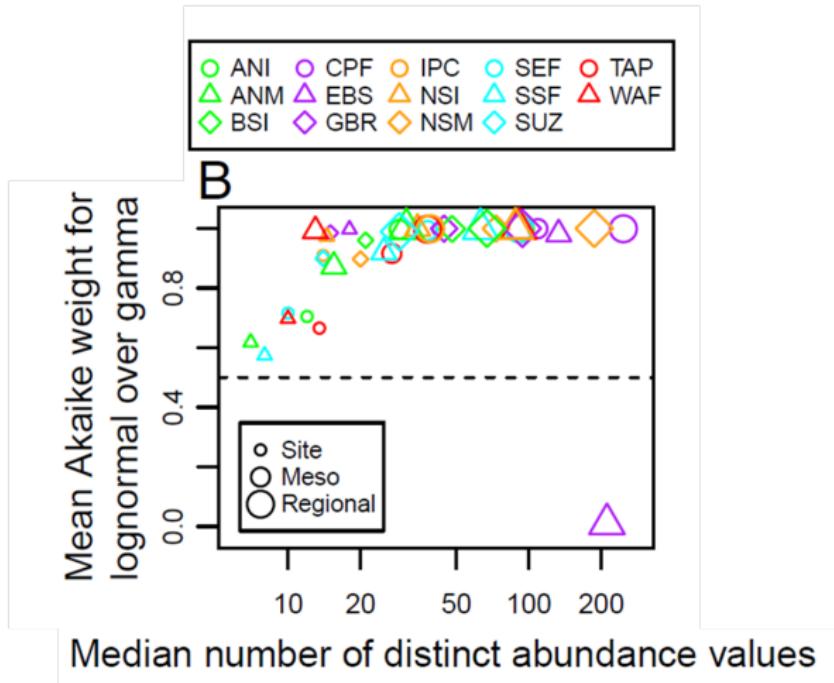
- Describes communities generated by neutral processes.

# NEUTRAL ANALYSIS

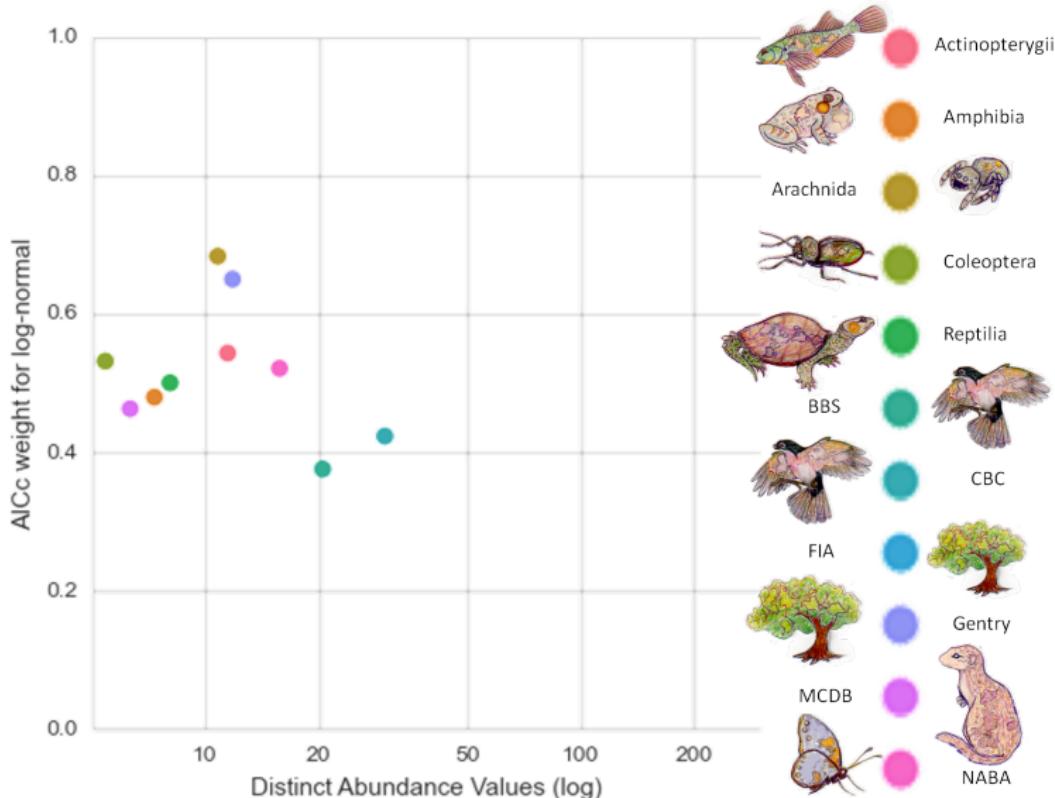


Poisson lognormal  
Negative binomial

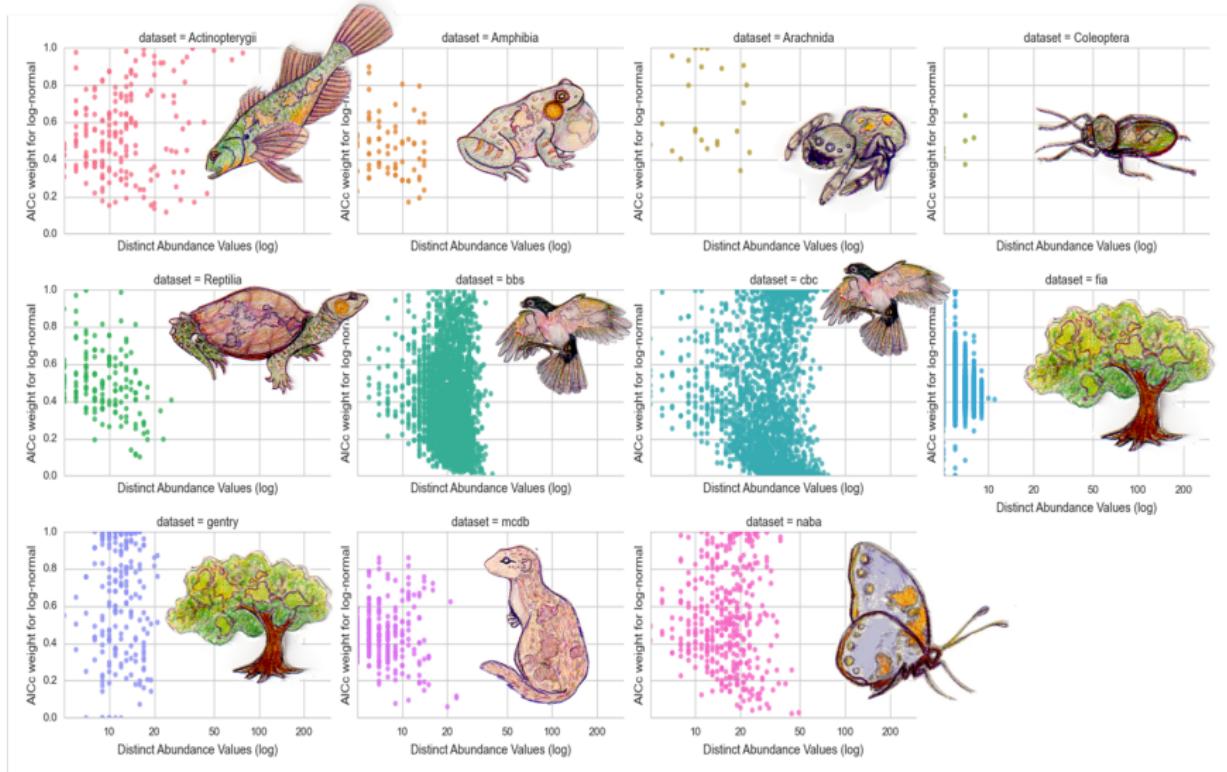
# NEUTRAL ANALYSIS



# NEUTRAL ANALYSIS



# NEUTRAL ANALYSIS



## NEUTRAL ANALYSIS

Difficult to identify a clear winning model.

- Demonstrates the importance of testing with multiple ecosystems.

## A DATA-INTENSIVE APPROACH

Challenging to infer process from species abundance distributions alone.

- Broad model categorization (i.e. neutral or non-neutral) may be more productive.
- May not be one single suite of processes that dominates.

## A DATA-INTENSIVE APPROACH

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.

## 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

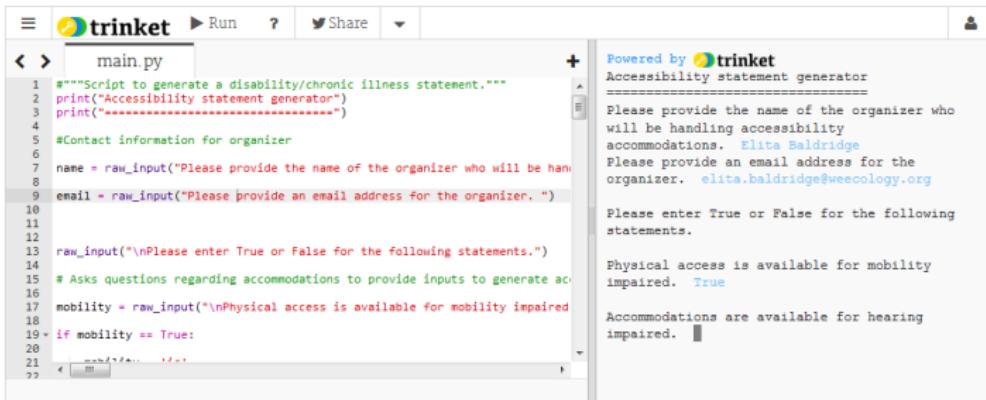
Dr. Thomas Price & USU Student Health Center.

A very supportive husband & family.

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

# THIS DISSERTATION BROUGHT TO YOU BY:

## Disability accommodations Tea, heating pads, & the Flint Hills of Kansas. Accessibility statement generator



The screenshot shows a Trinket-based Python code editor. The code in `main.py` is as follows:

```
1 """Script to generate a disability/chronic illness statement."""
2 print("Accessibility statement generator")
3 print("-----")
4
5 #Contact information for organizer
6
7 name = raw_input("Please provide the name of the organizer who will be handling accessibility accommodations. ")
8 email = raw_input("Please provide an email address for the organizer. ")
9
10
11
12 raw_input("\nPlease enter True or False for the following statements.")
13
14 # Asks questions regarding accommodations to provide inputs to generate accessibility statement
15 mobility = raw_input("\nPhysical access is available for mobility impaired: ")
16
17 if mobility == True:
18     mobility = "True"
19 else:
20     mobility = "False"
21
22
```

The right panel shows the generated accessibility statement:

Powered by trinket  
Accessibility statement generator  
-----  
Please provide the name of the organizer who will be handling accessibility accommodations. Eliza Baldridge  
Please provide an email address for the organizer. eliza.baldridge@weecology.org  
  
Please enter True or False for the following statements.  
  
Physical access is available for mobility impaired. True  
  
Accommodations are available for hearing impaired.

<https://trinket.io/python/2df3ea45cf>

# QUESTIONS?



# ABUNDANCE DATABASE

