A data-intensive assessment of the species-abundance distribution.

Elita Baldridge
@elitabaldridge

Utah State University
ECOLOGY CENTER

Feel free to:

Copy, share, adapt, or re-mix;

Photograph, film, or broadcast;

Blog, live-blog, or post video of;

Provided that:

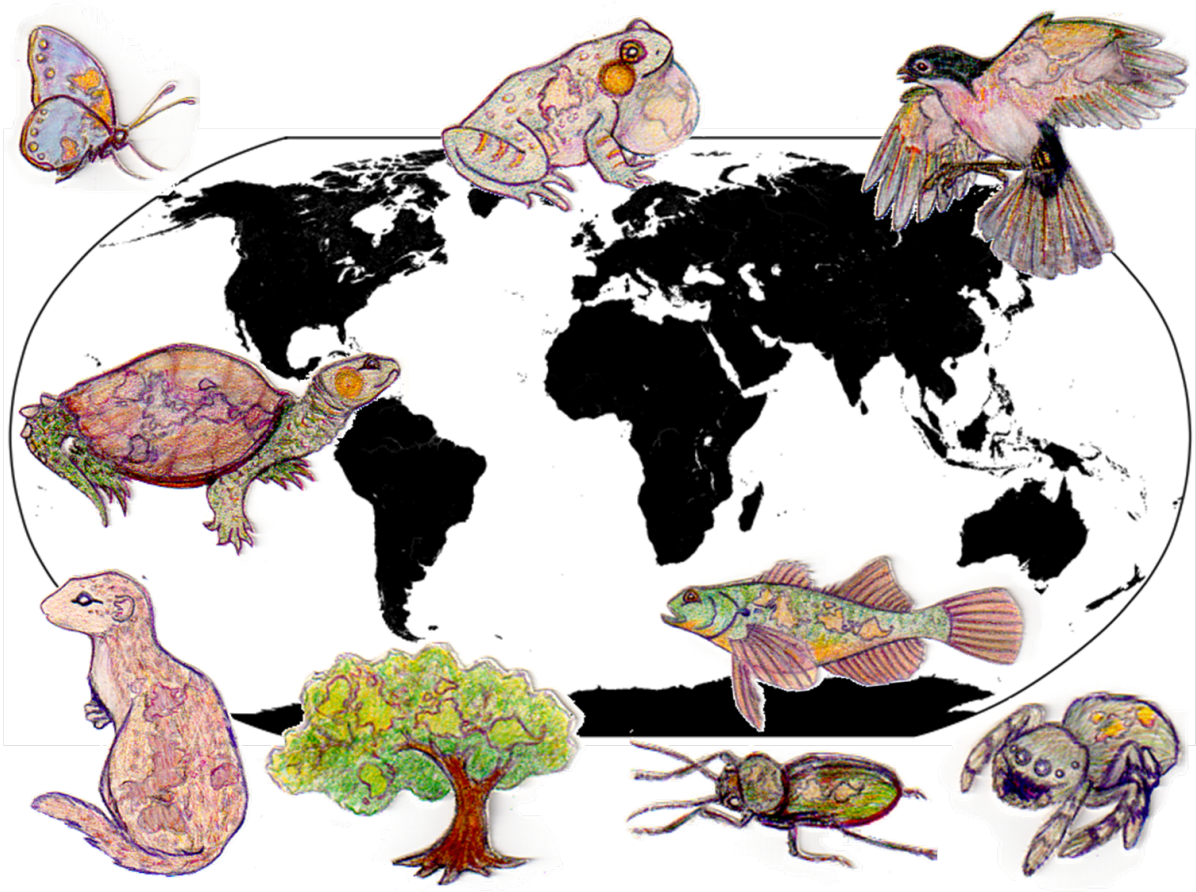
You attribute the work to its author and respect the rights and licenses associated with its components.

Open science

- Code:
  - github.com/embaldridge
  - github.com/weecology
- Data: figshare.com

Macroecology

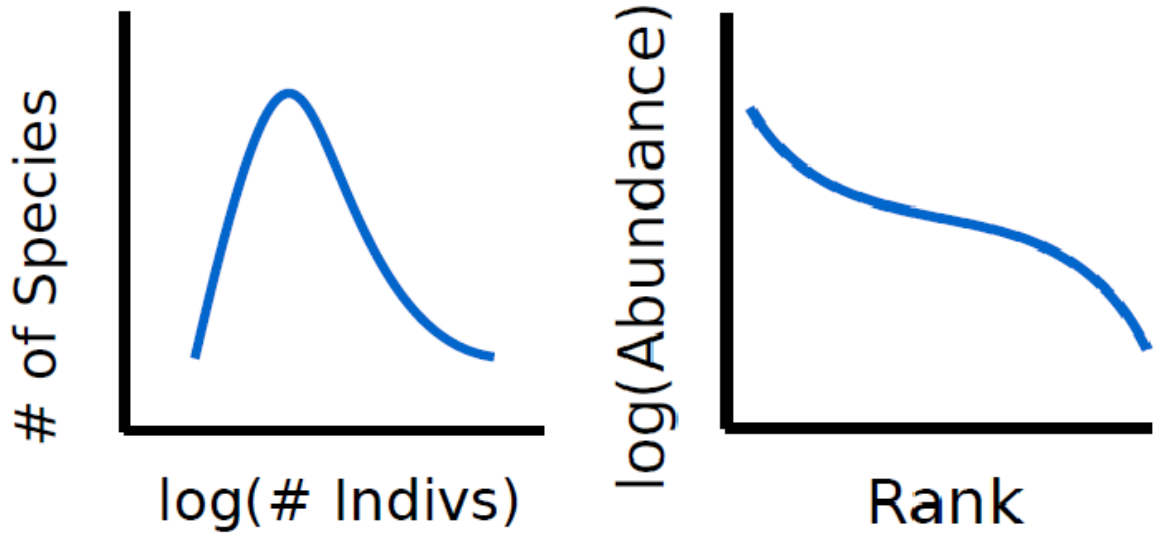
One approach for identifying general ecological patterns & processes.



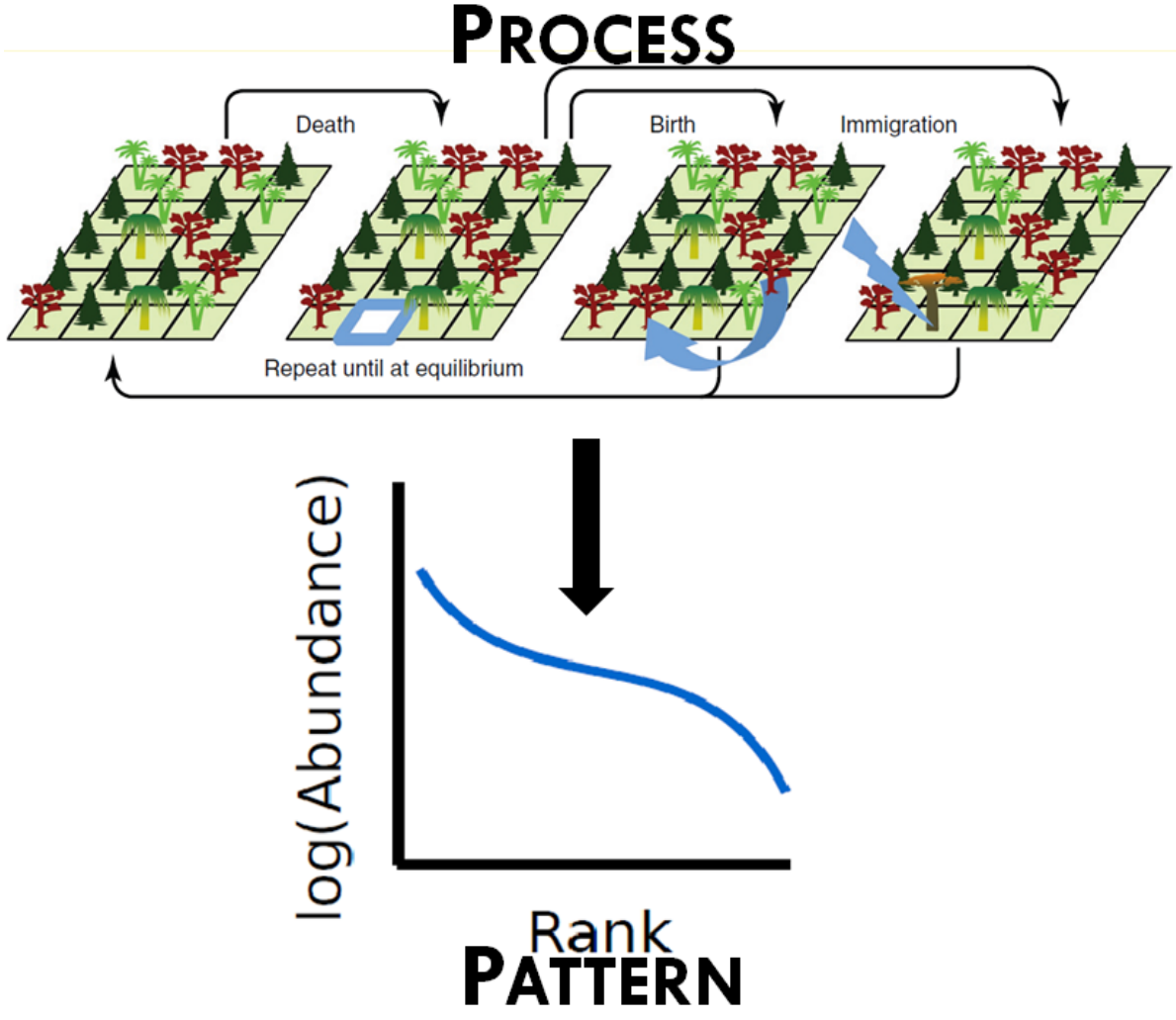
Let's begin with an example…

Species abundance distribution (SAD)

- Describes the distribution of commonness & rarity of species.



Let's begin with an example...

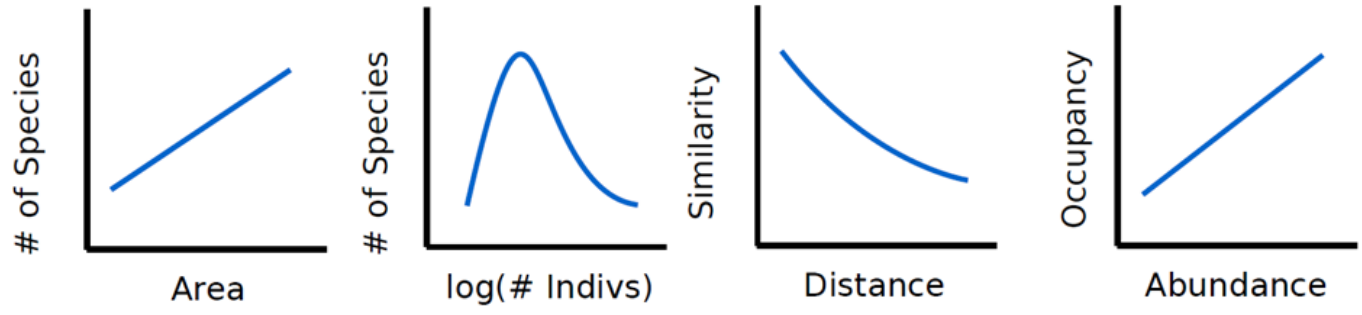


Patterns & process.

Process

Competition Niches Dispersal

Nutrients Neutrality Etc.



Pattern

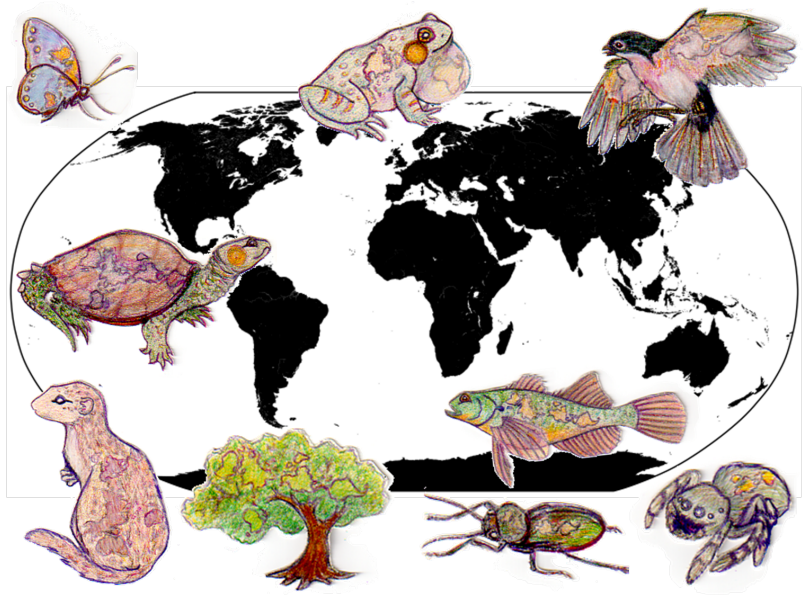
Pattern & Process; Signal & Noise

Challenges of macroecology

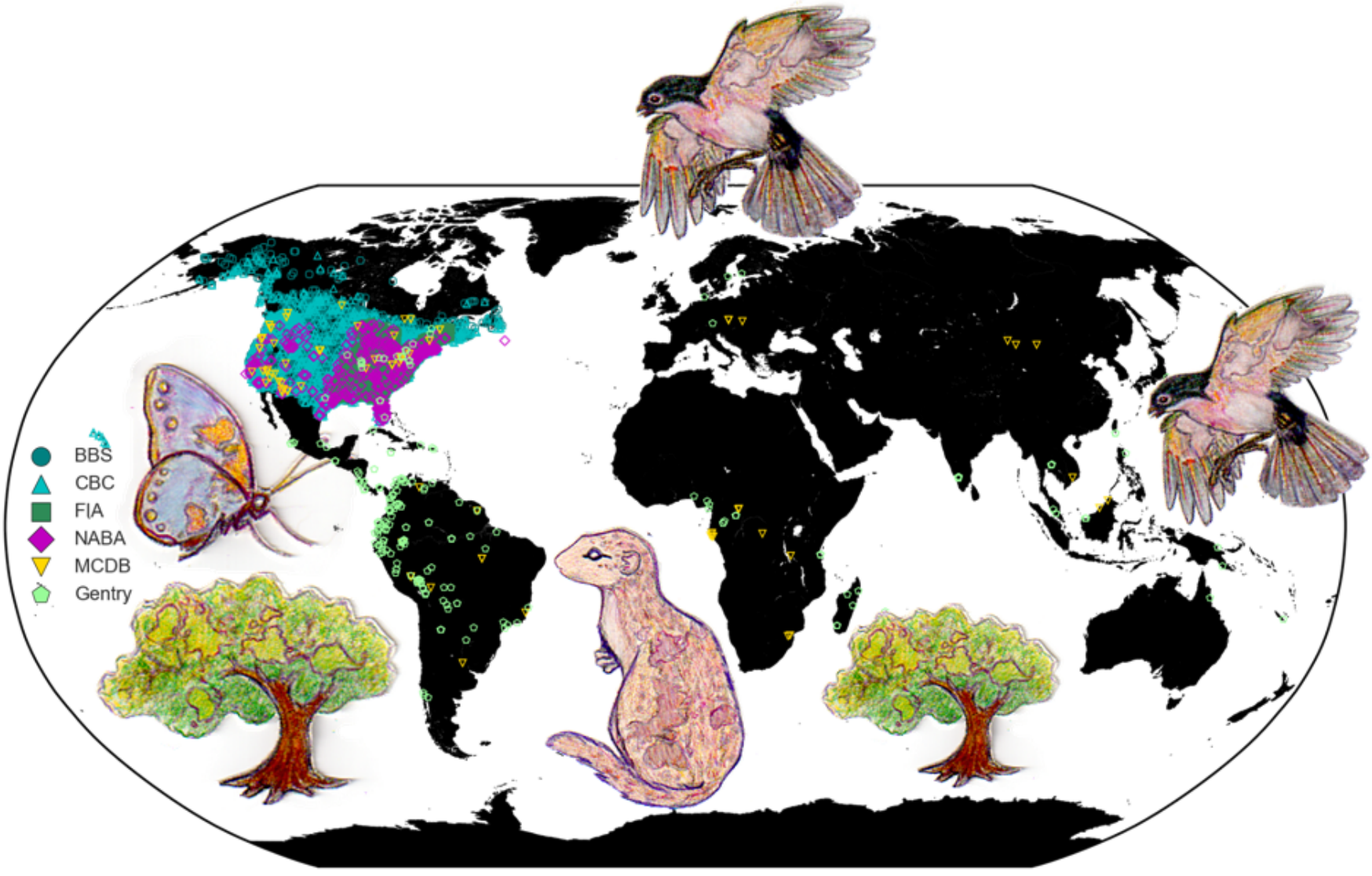
- Studies performed with a limited number of large datasets.
- Lack of identification of pattern generating mechanisms.

Best practice recommendations

- Test multiple models with consistent statistical approach.
- Test with multiple taxonomic groups/ecosystems.



Current Data



Macroecological data

Challenges of macroecology

- Limited number of large datasets.
- 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

- Decide on inclusion criteria.
- Decide what variables to collect.
- Decide on handling of missing data.
- Decide on database structure.
- Search for data.
- Sort papers into useable and not.
- Collect data.

Record decisions at all steps:

Metadata are important.

Abundance database

Inclusion criteria:

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

Abundance database

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 wrangling

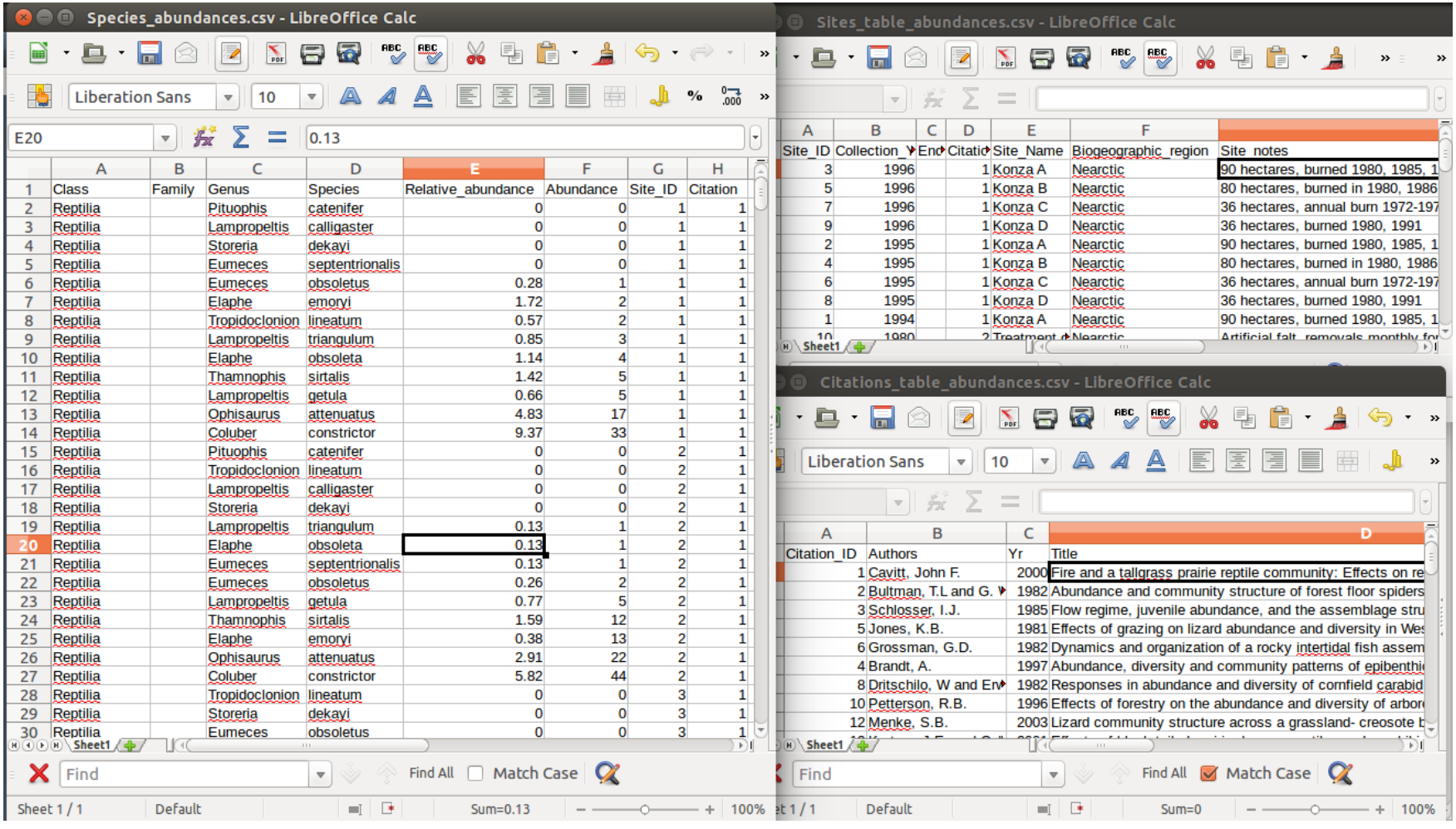
– Fire and a Tallgrass Prairie Reptile Community: Effects on Relative Abundance and Seasonal Activity

p. 14 (4 of 10)

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	
Number of Trap arrays	4	9	9	9	9	4	4	4	4	Total
Snakes										
**Coluber constrictor*	33	44	52	58	20	4	31	15	28	285
	9.37	5.82	6.64	7.67	2.55	1.82	7.75	7.81	7.00	
**Thamnophis sirtalis*	5	12	9	15	10	21	10	26	7	115
	1.42	1.59	1.15	1.98	1.28	9.55	2.50	13.54	1.75	
Elaphe emoryi	2	13	3	3	7	5	3	8	4	48
	0.57	1.72	0.38	0.40	0.89	2.27	0.75	4.17	1.0	
Lampropeltis getula	5	5	6	7	7	1	1	1	5	38
	1.42	0.66	0.77	0.93	0.89	0.45	0.25	0.52	1.25	
Lampropeltis triangulum	3	1	3	7	4	—	1	3	4	26
	0.85	0.13	0.38	0.93	0.51	—	0.25	1.56	1.0	
Pituophis catenifer	—	—	2	3	—	2	2	5	12	26
	—	—	0.26	0.40	—	0.91	0.5	2.60	3.0	
Elaphe obsoleta	4	1	1	—	2	—	—	—	—	8
	1.14	0.13	0.13	—	0.26	—	—	—	—	
Tropidoclonion lineatum	2	—	—	—	—	—	—	—	—	2
	0.57	—	—	—	—	—	—	—	—	
Lampropeltis calligaster	—	—	1	—	—	—	—	—	—	1
	—	—	0.13	—	—	—	—	—	—	
Storeria dekayi	—	—	—	—	—	—	—	1	—	1
								0.52		
Lizards										
**Ophisaurus attenuatus*	17	22	10	30	10	2	—	—	—	91
	4.83	2.91	1.28	3.97	1.28	0.91	—	—	—	
Eumeces obsoletus	1	2	—	6	1	—	1	—	—	11
	0.28	0.26	—	0.79	0.13	—	0.25	—	—	
Eumeces septentrionalis	—	1	—	5	—	—	—	—	—	6
	—	0.13	—	0.66	—	—	—	—	—	

Data wrangling

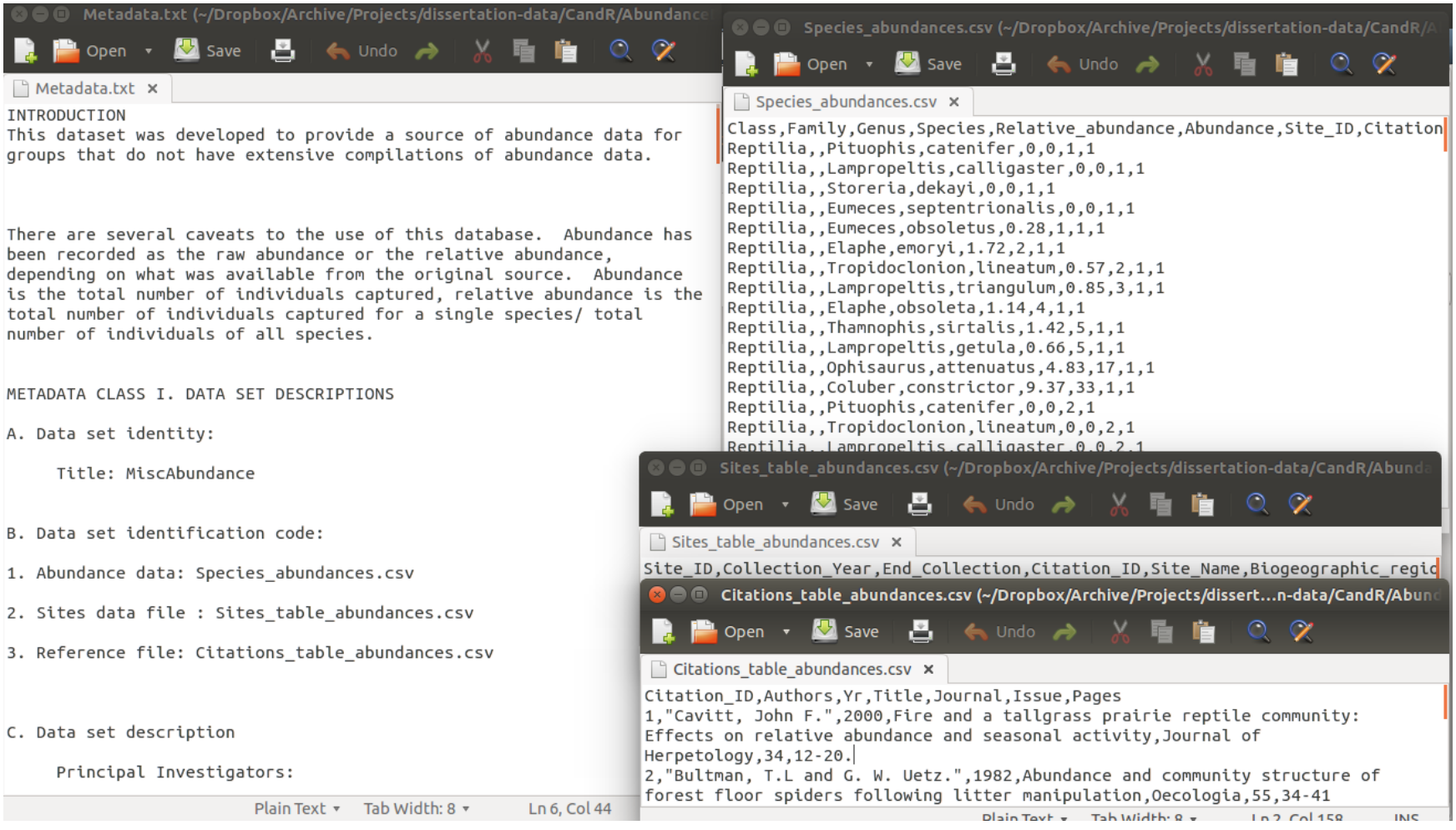


	A	B	C	D	E	F	G	H
1	Class	Family	Genus	Species	Relative_abundance	Abundance	Site_ID	Citation
2	Reptilia		Pituophis	catenifer	0	0	1	1
3	Reptilia		Lampropeltis	calligaster	0	0	1	1
4	Reptilia		Storeria	dekayi	0	0	1	1
5	Reptilia		Eumeces	septentrionalis	0	0	1	1
6	Reptilia		Eumeces	obsoletus	0.28	1	1	1
7	Reptilia		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		Pituophis	catenifer	0	0	2	1
16	Reptilia		Tropidoclonion	lineatum	0	0	2	1
17	Reptilia		Lampropeltis	calligaster	0	0	2	1
18	Reptilia		Storeria	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	obsoletus	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		Storeria	dekayi	0	0	3	1
30	Reptilia		Eumeces	obsoletus	0	0	3	1

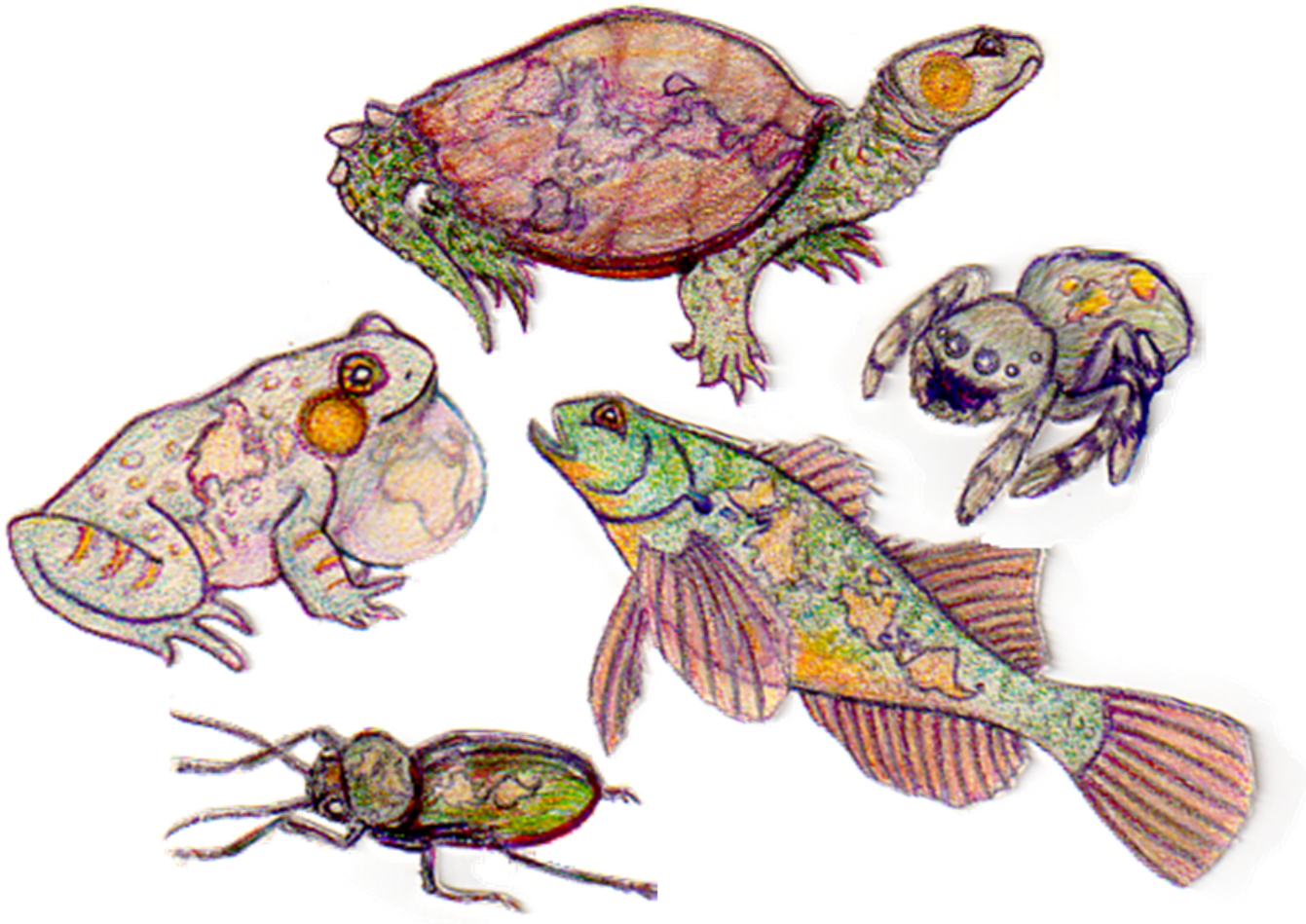
	A	B	C	D	E	F
1	Site_ID	Collection_Y	End	Citatio	Site_Name	Biogeographic_region	Site_notes
2	3	1996		1	Konza A	Nearctic	90 hectares, burned 1980, 1985, 1
3	5	1996		1	Konza B	Nearctic	80 hectares, burned in 1980, 1986
4	7	1996		1	Konza C	Nearctic	36 hectares, annual burn 1972-197
5	9	1996		1	Konza D	Nearctic	36 hectares, burned 1980, 1991
6	2	1995		1	Konza A	Nearctic	90 hectares, burned 1980, 1985, 1
7	4	1995		1	Konza B	Nearctic	80 hectares, burned in 1980, 1986
8	6	1995		1	Konza C	Nearctic	36 hectares, annual burn 1972-197
9	8	1995		1	Konza D	Nearctic	36 hectares, burned 1980, 1991
10	1	1994		1	Konza A	Nearctic	90 hectares, burned 1980, 1985, 1
11	10	1980		2	Treatment A	Nearctic	Artificial fall removals monthly for

	A	B	C	D
1	Citation_ID	Authors	Yr	Title
2	1	Cavitt, John F.	2000	Fire and a tallgrass prairie reptile community: Effects on re
3	2	Bultman, T.L and G. W	1982	Abundance and community structure of forest floor spiders
4	3	Schlosser, I.J.	1985	Flow regime, juvenile abundance, and the assemblage stru
5	5	Jones, K.B.	1981	Effects of grazing on lizard abundance and diversity in Wes
6	6	Grossman, G.D.	1982	Dynamics and organization of a rocky intertidal fish assem
7	4	Brandt, A.	1997	Abundance, diversity and community patterns of epibenthic
8	8	Dritschilo, W and En	1982	Responses in abundance and diversity of cornfield carabid
9	10	Petterson, R.B.	1996	Effects of forestry on the abundance and diversity of arbor
10	12	Menke, S.B.	2003	Lizard community structure across a grassland- creosote b

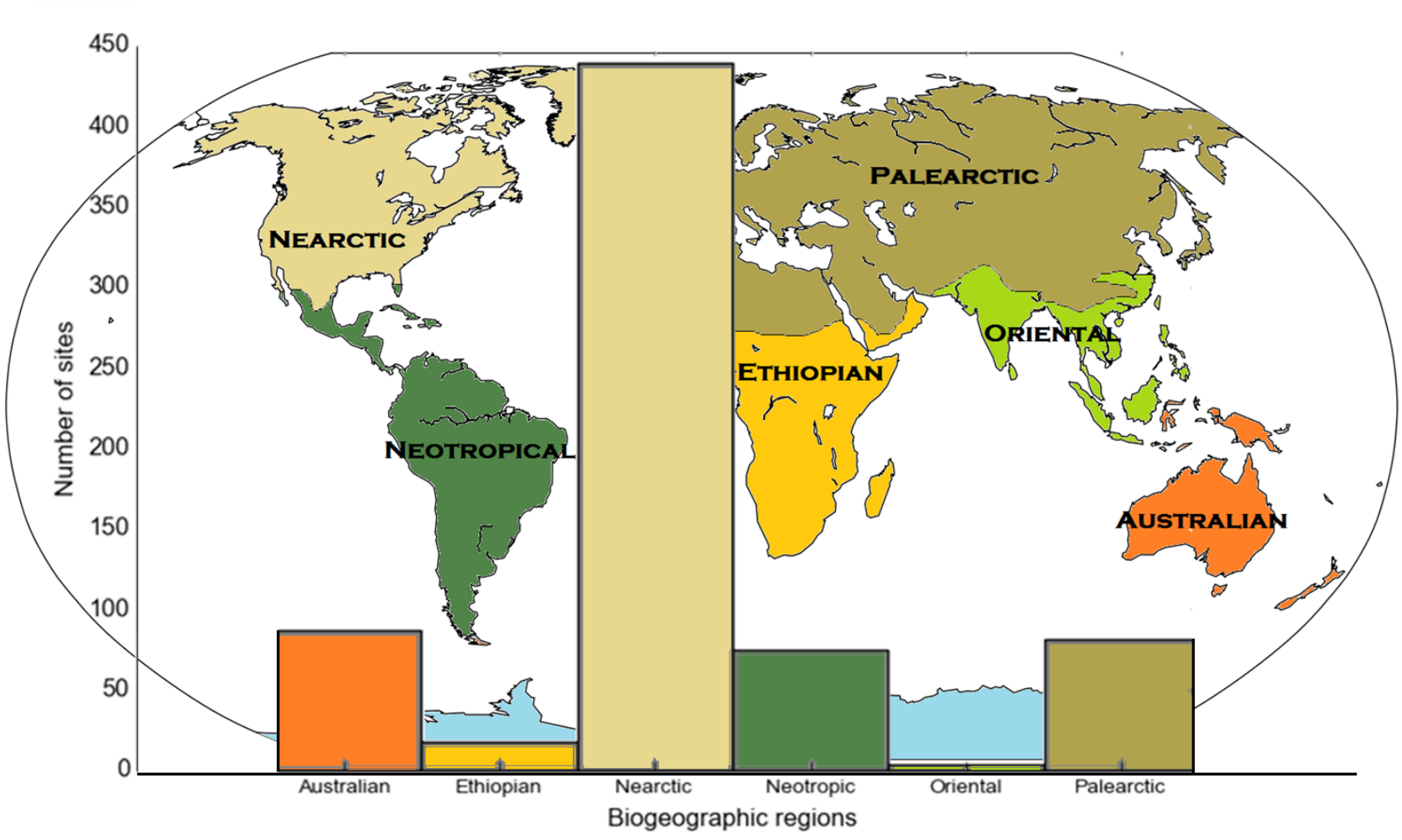
Data wrangling



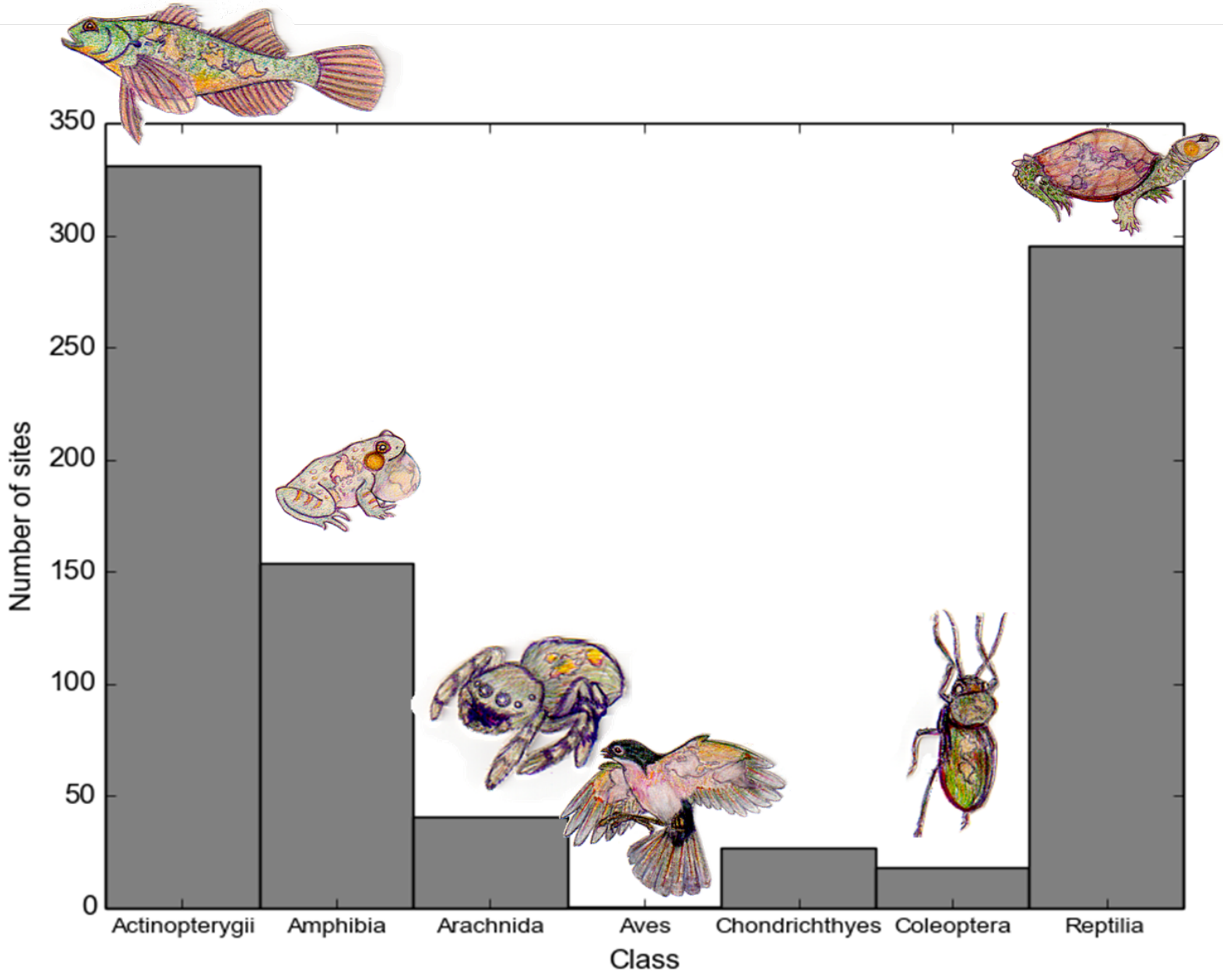
Abundance database



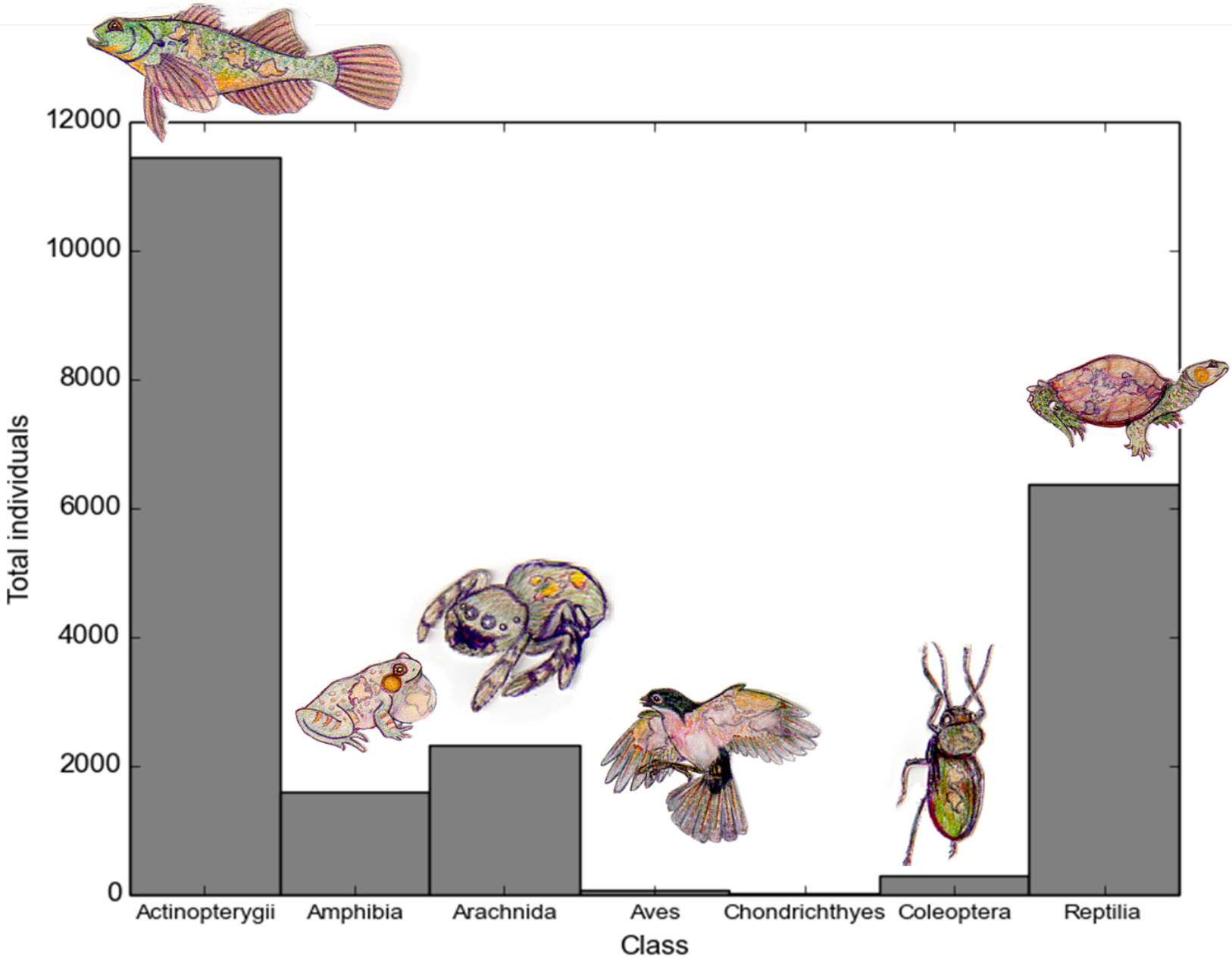
Abundance database



Abundance database



Abundance database



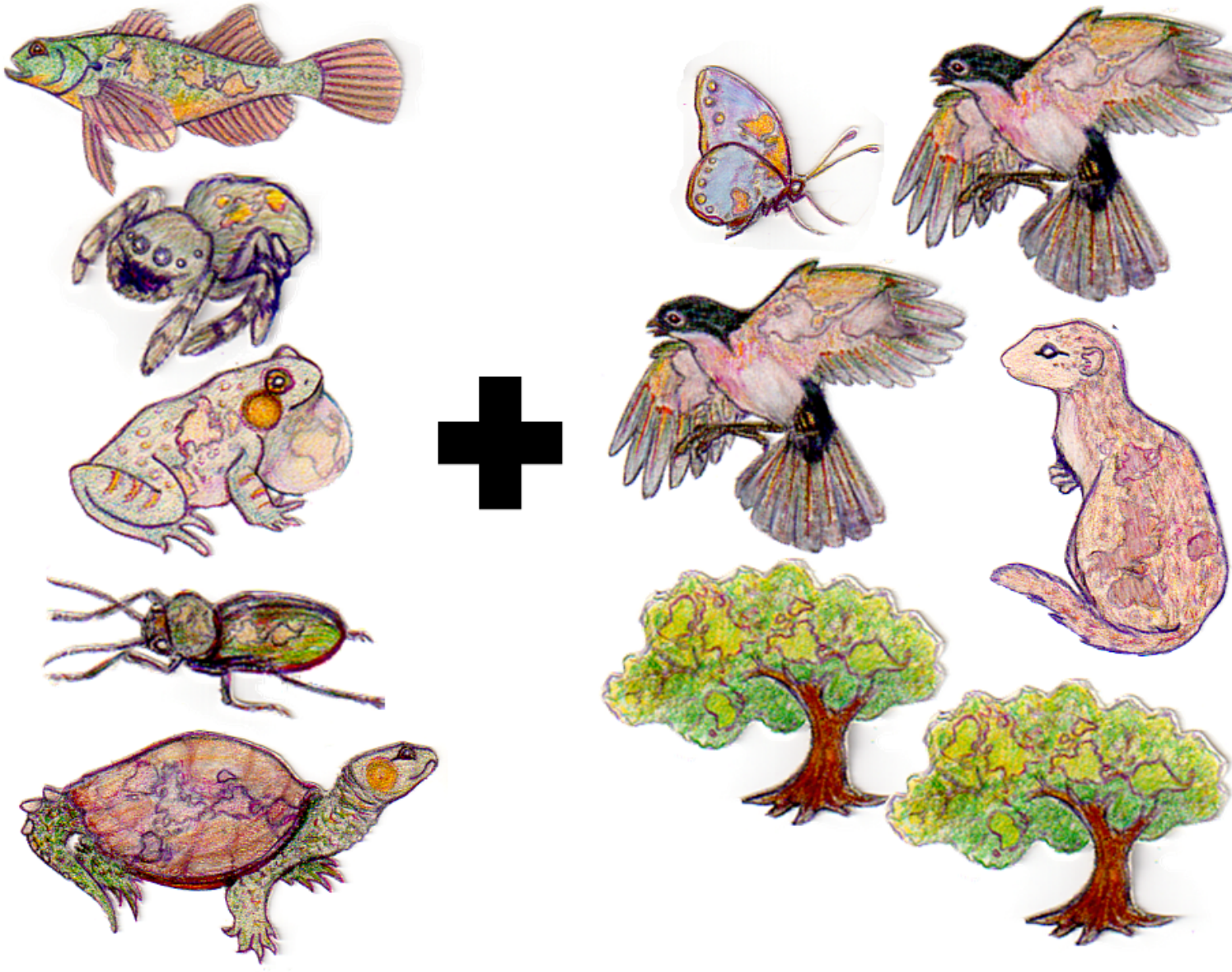
Data Availability

Public & open access through figshare.
EcoData Retriever importable.

```
(http://figshare.com)
(http://www.ecodataretriever.org)
sad_data =
ecoretriever::fetch('MiscAbundanceDB')
```

Data

Data



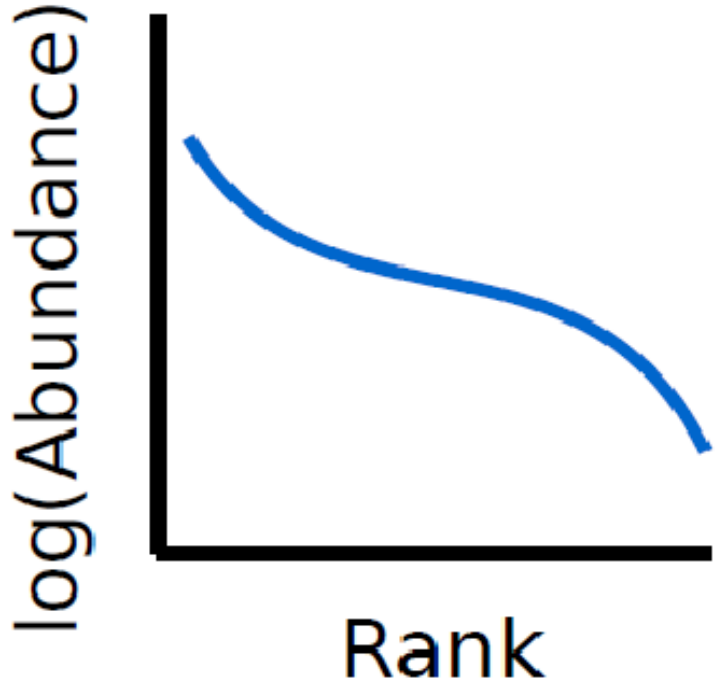
SAD Comparisons

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

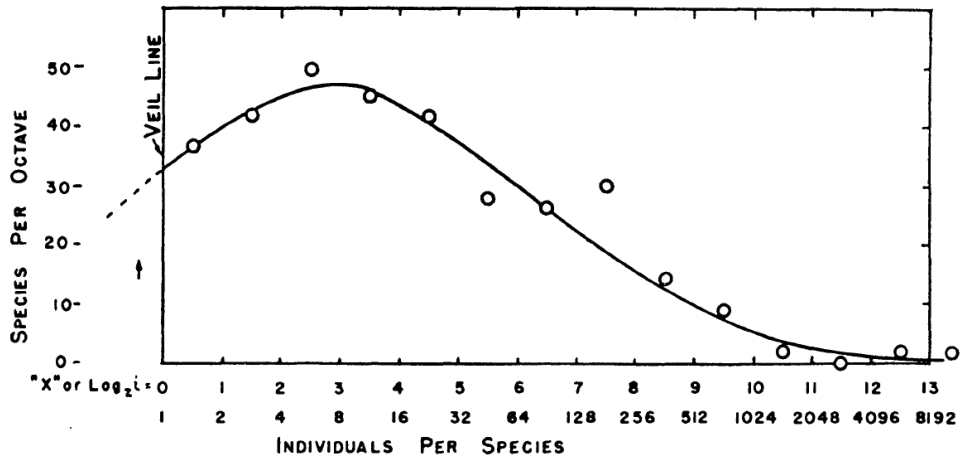
The species abundance distribution:



Many models of the species abundance distribution (SAD).

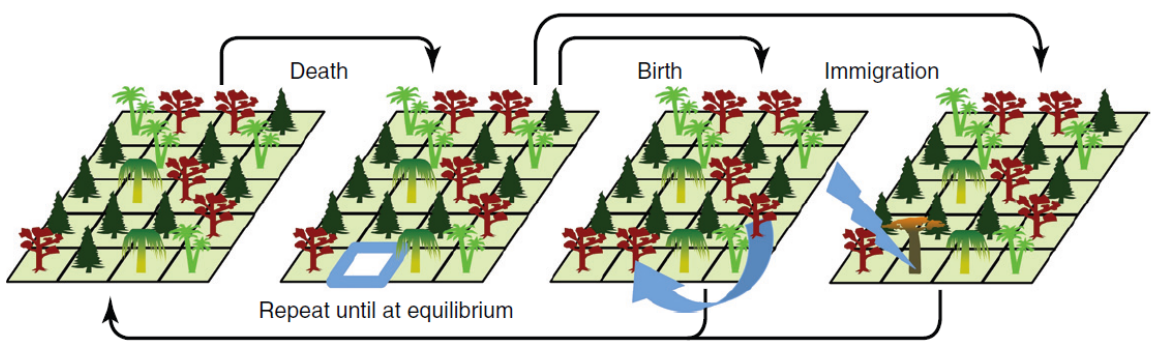
SAD models

Statistical description



Preston 1962a.

Process-based



Rosindell et al. 2011.

SAD Comparisons

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?
- The model with the greatest AICc weight was the ”winner”.

SAD Comparisons

Computational tools:

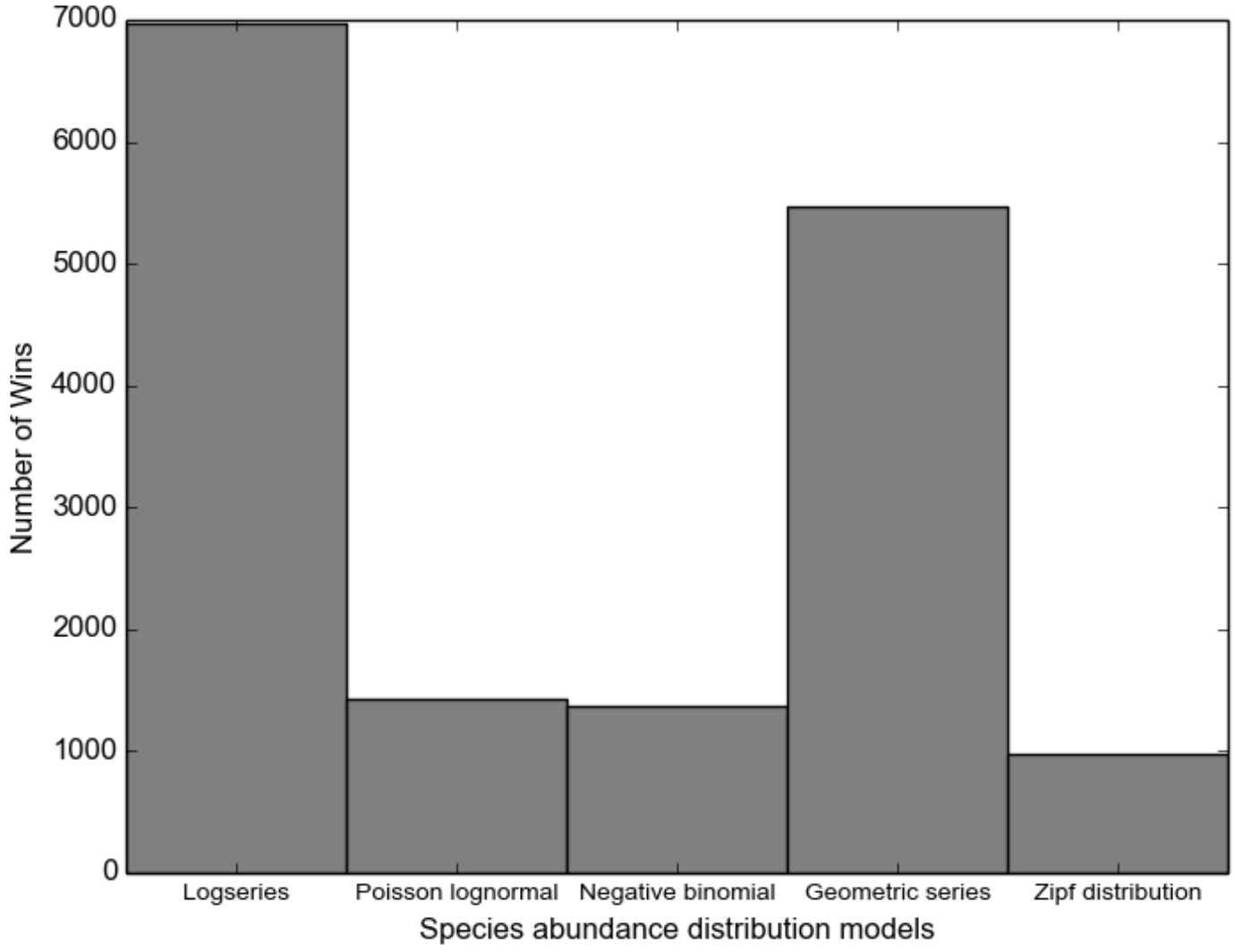
- Model fitting, log-likelihood, & AICc: macroecotools Python package.
  (https://github.com/weecology/macroecotools)
- All code & majority of data are publicly available.
  (https://github.com/weecology/sad-comparison)

Open science

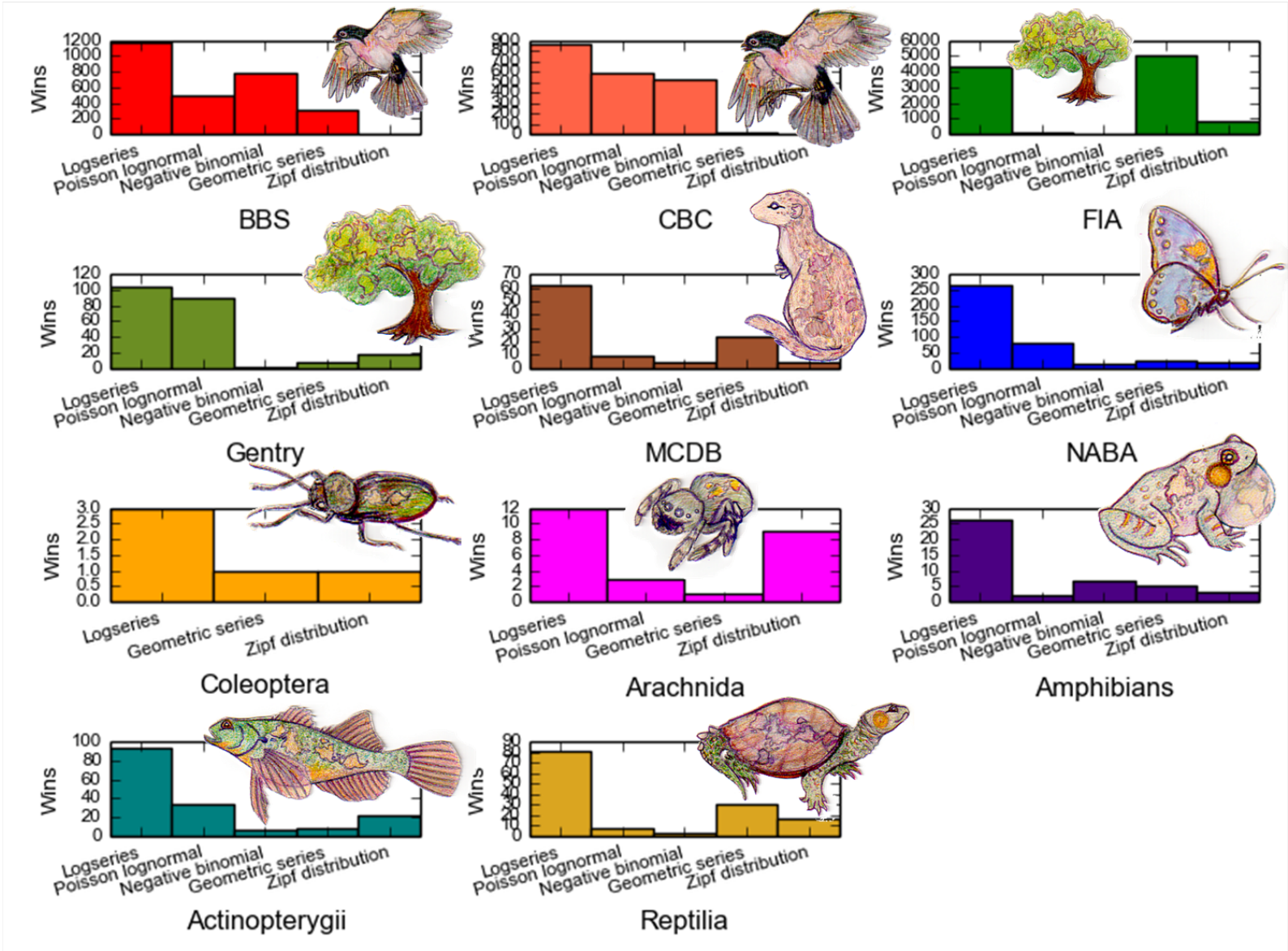
More reproducible.
More accessible.

Better science.

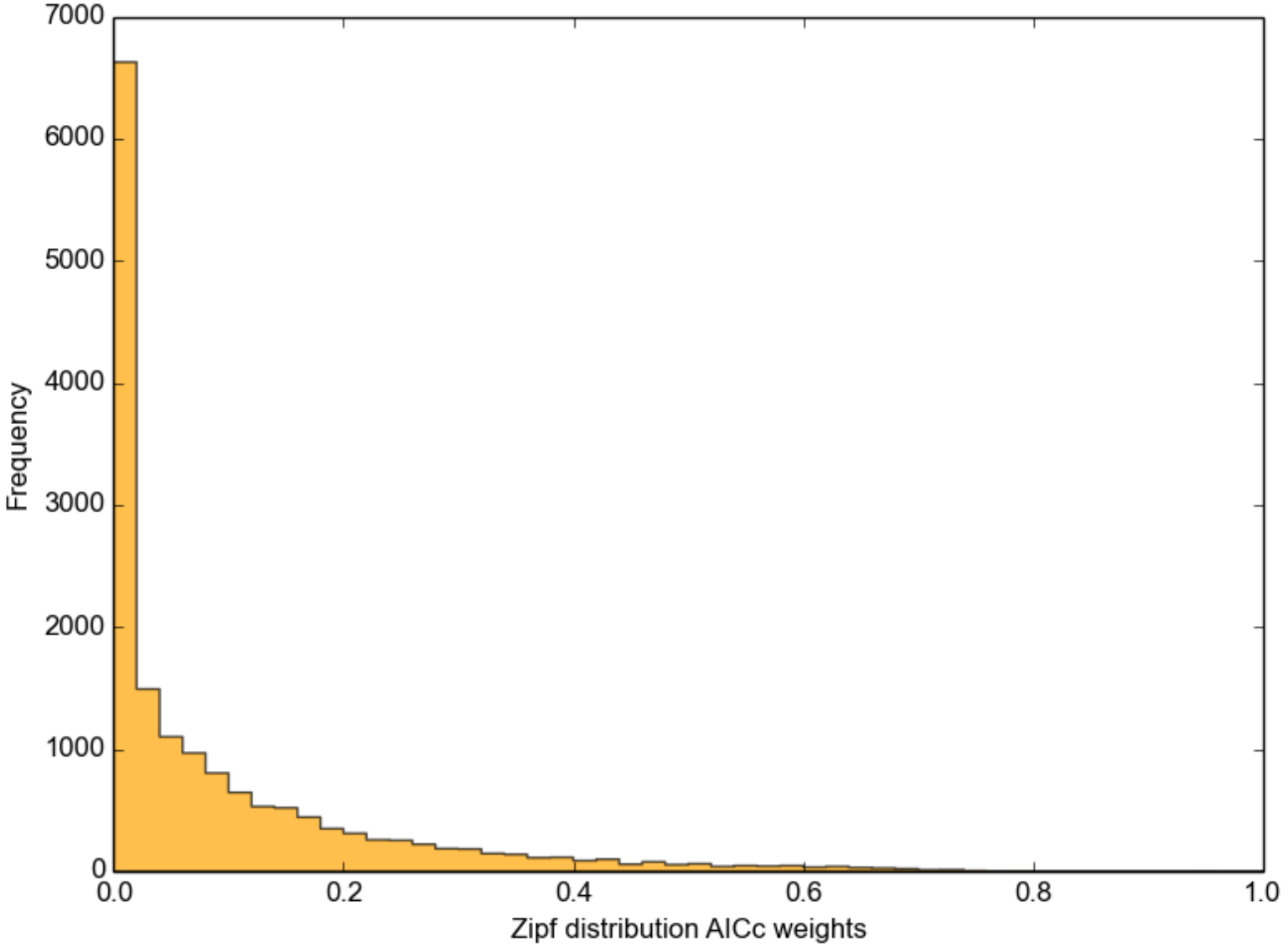
SAD Comparisons



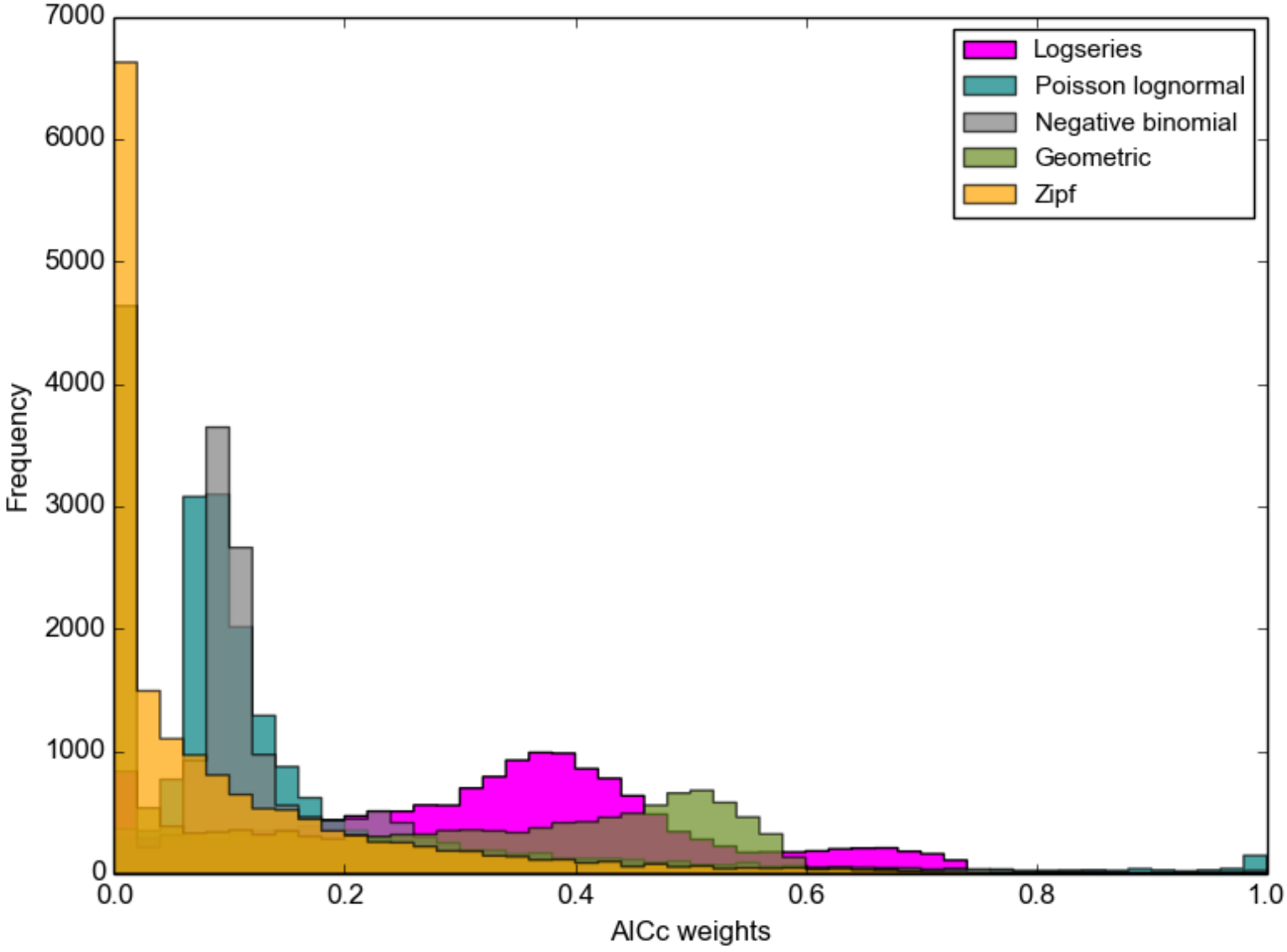
SAD Comparisons



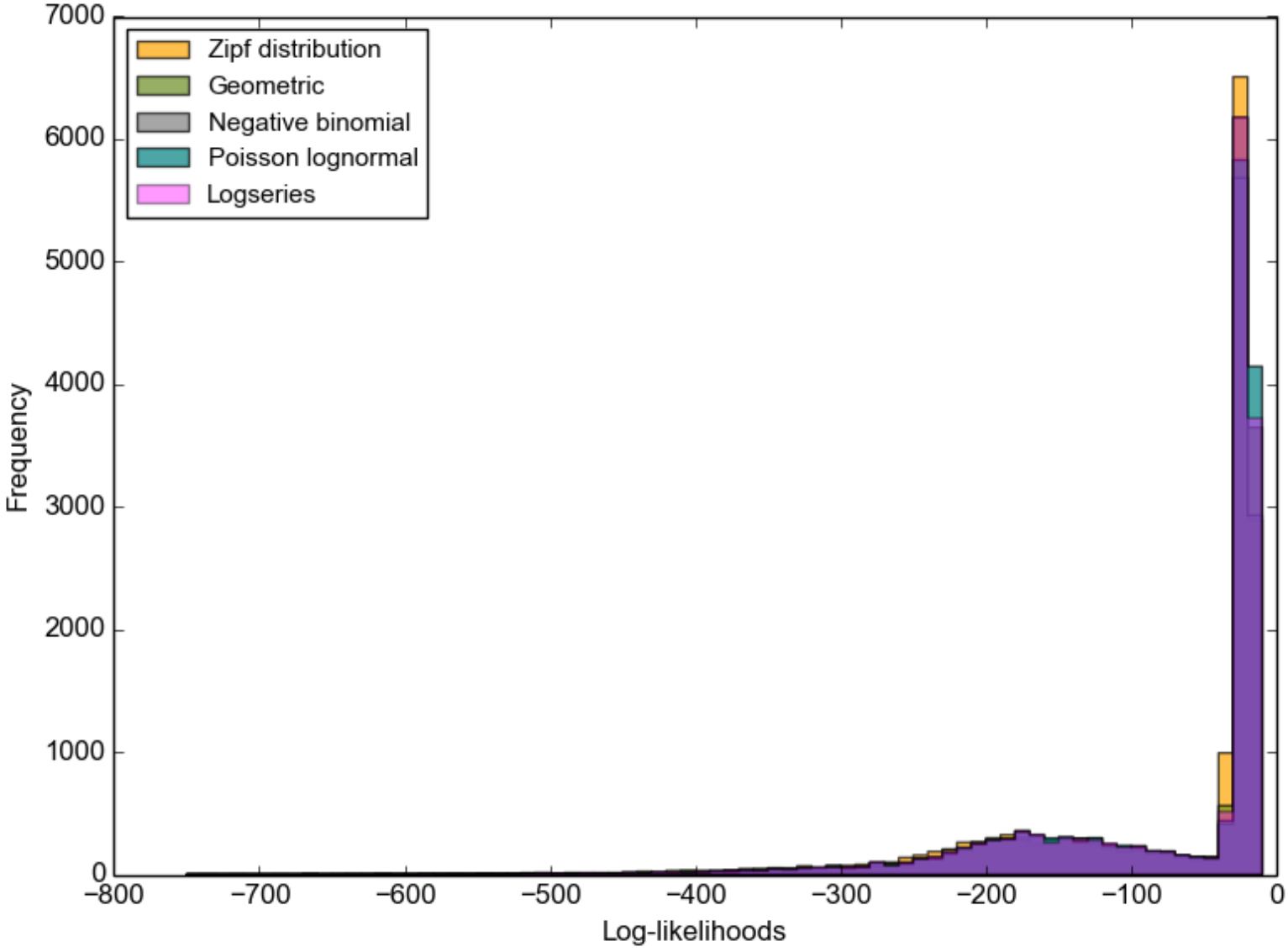
SAD Comparisons



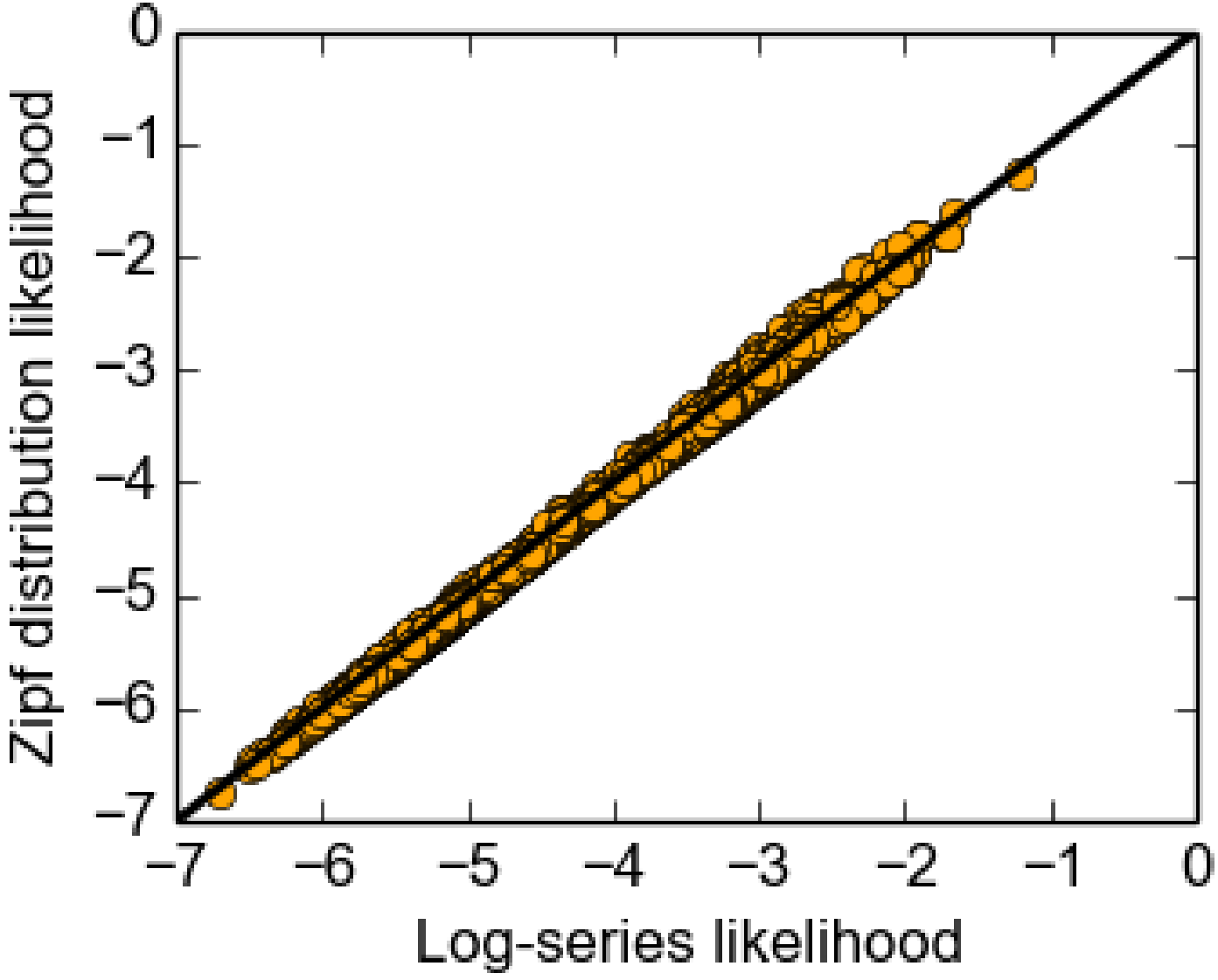
SAD Comparisons



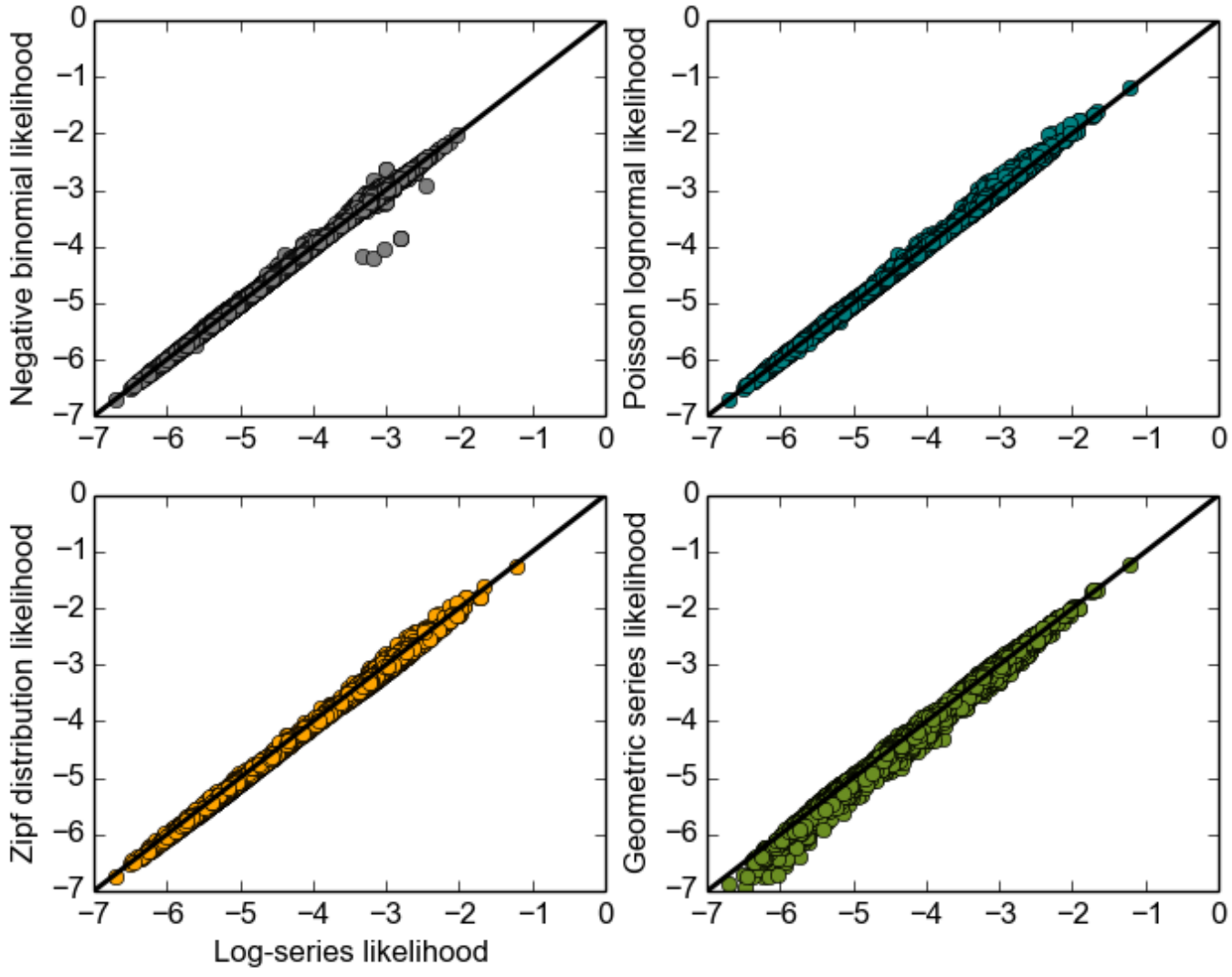
SAD Comparisons



SAD Comparisons



SAD Comparisons



SAD Comparisons

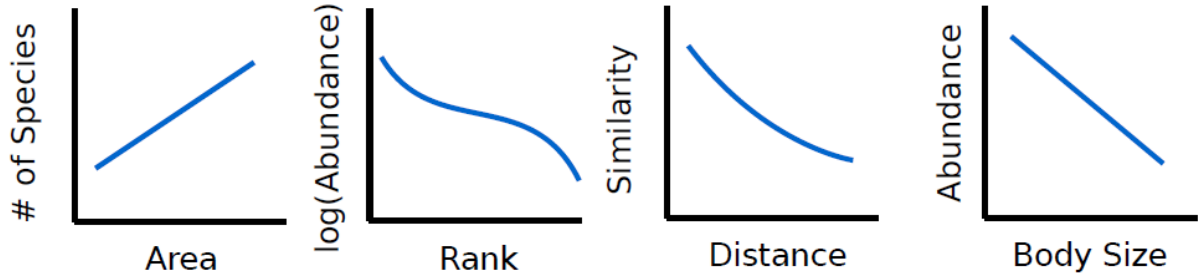
Existing models provide equivalently good absolute fits to empirical data.

- 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.



SAD Comparisons

Identifying pattern generating mechanisms:

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

Neutral Analysis

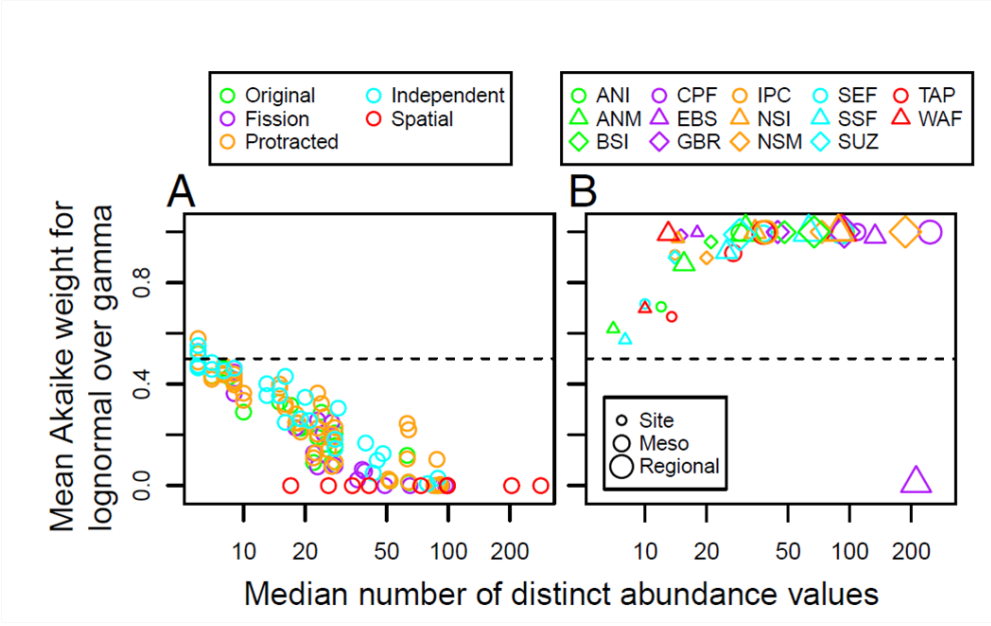
Early tests of neutral theory compared the fit of empirical species abundance distributions to the neutral prediction.

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.

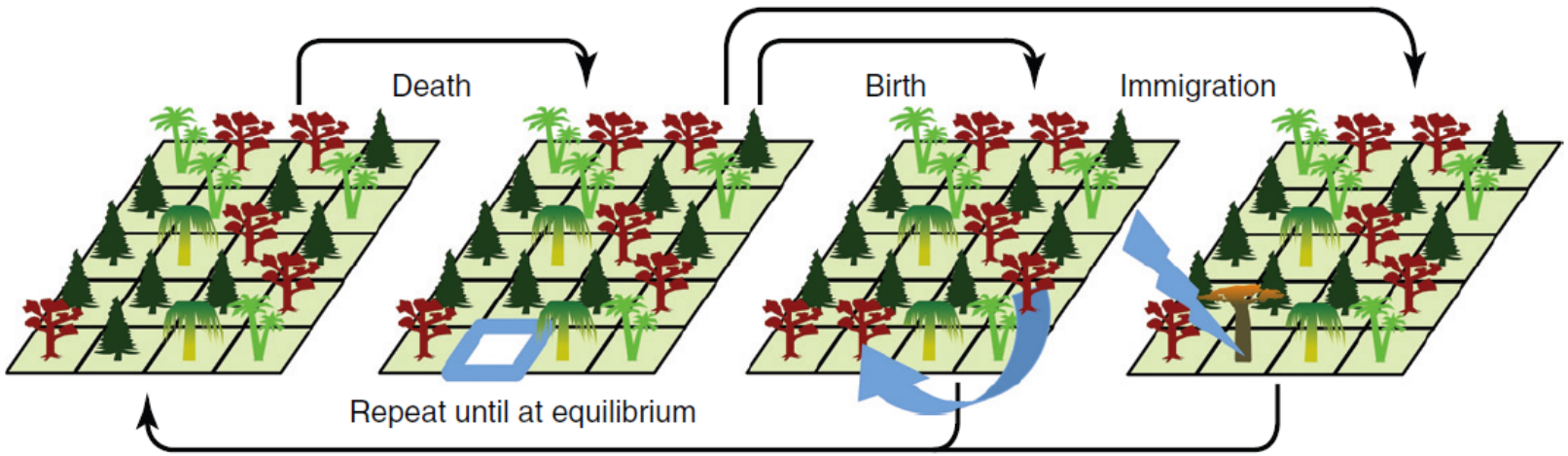


May be a robust method for identifying communities that exhibit non-neutrality.

Neutral theory

Many formulations, but:

- Species & individuals ecologically & demographically equivalent.

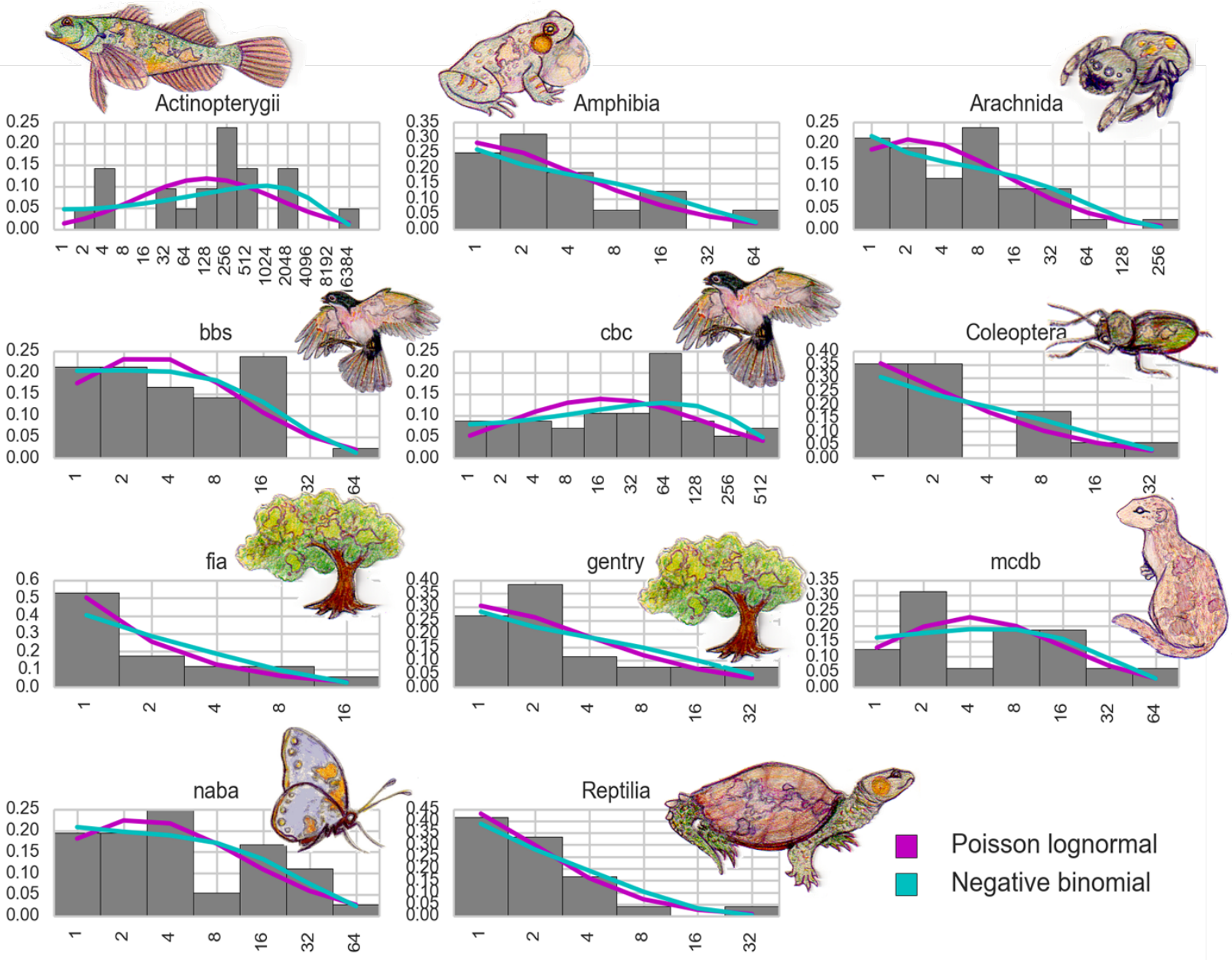


Rosindell et al. 2011.

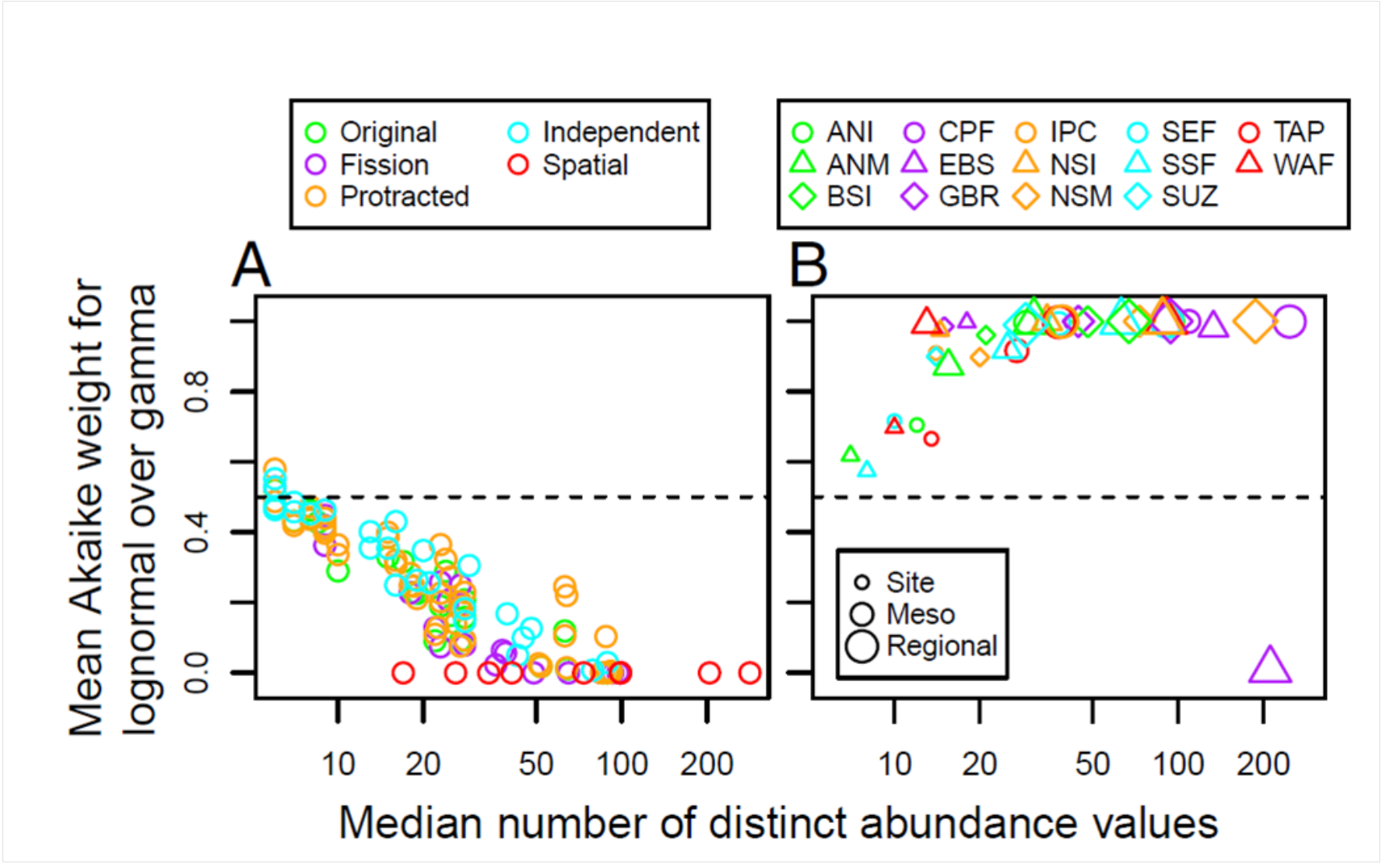
Neutral Analysis

Used the same data and model fitting approach.
Compared a non-neutral model (Poisson lognormal) to a neutral model (negative binomial).

Neutral Analysis

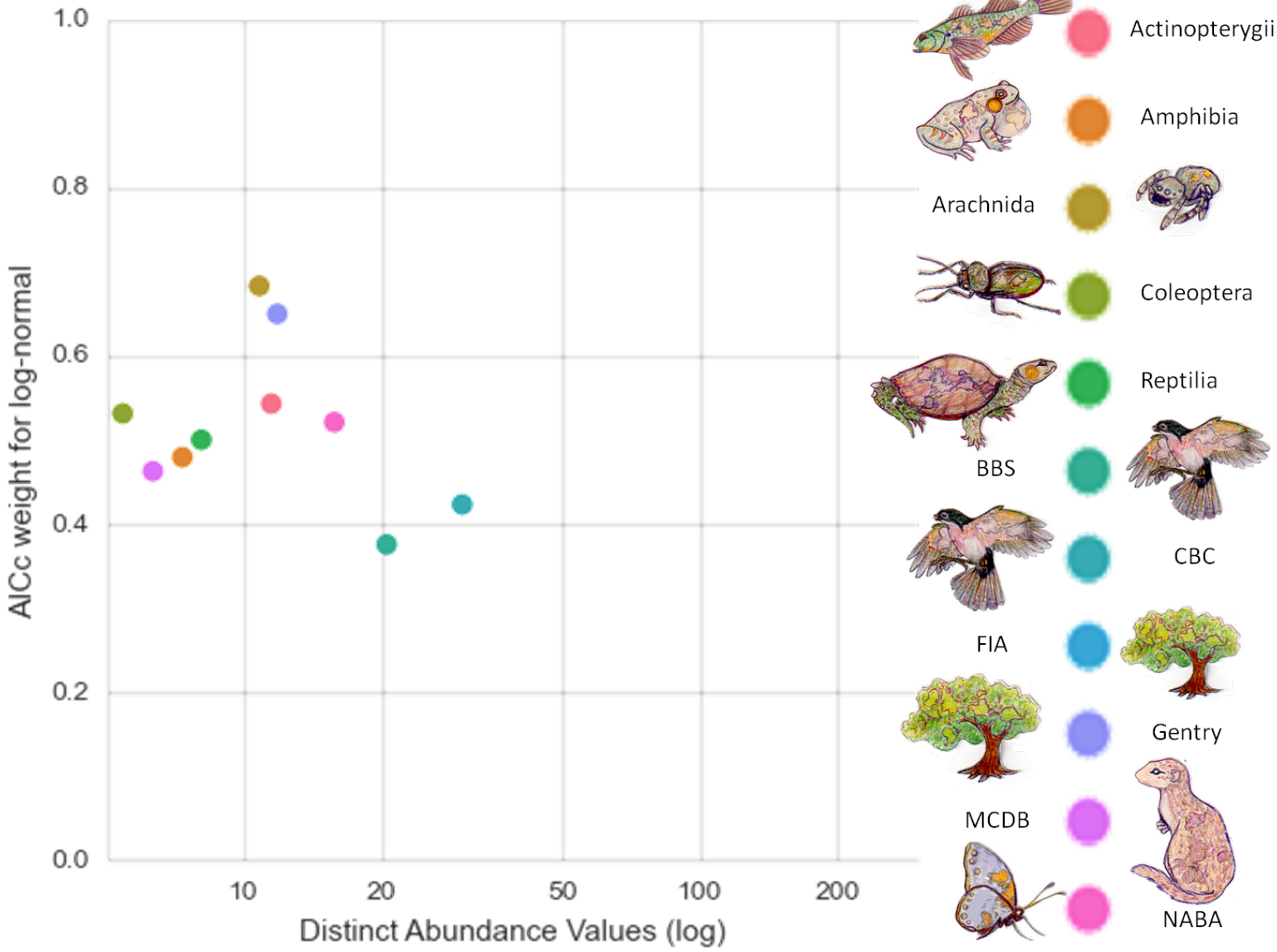


Neutral Analysis

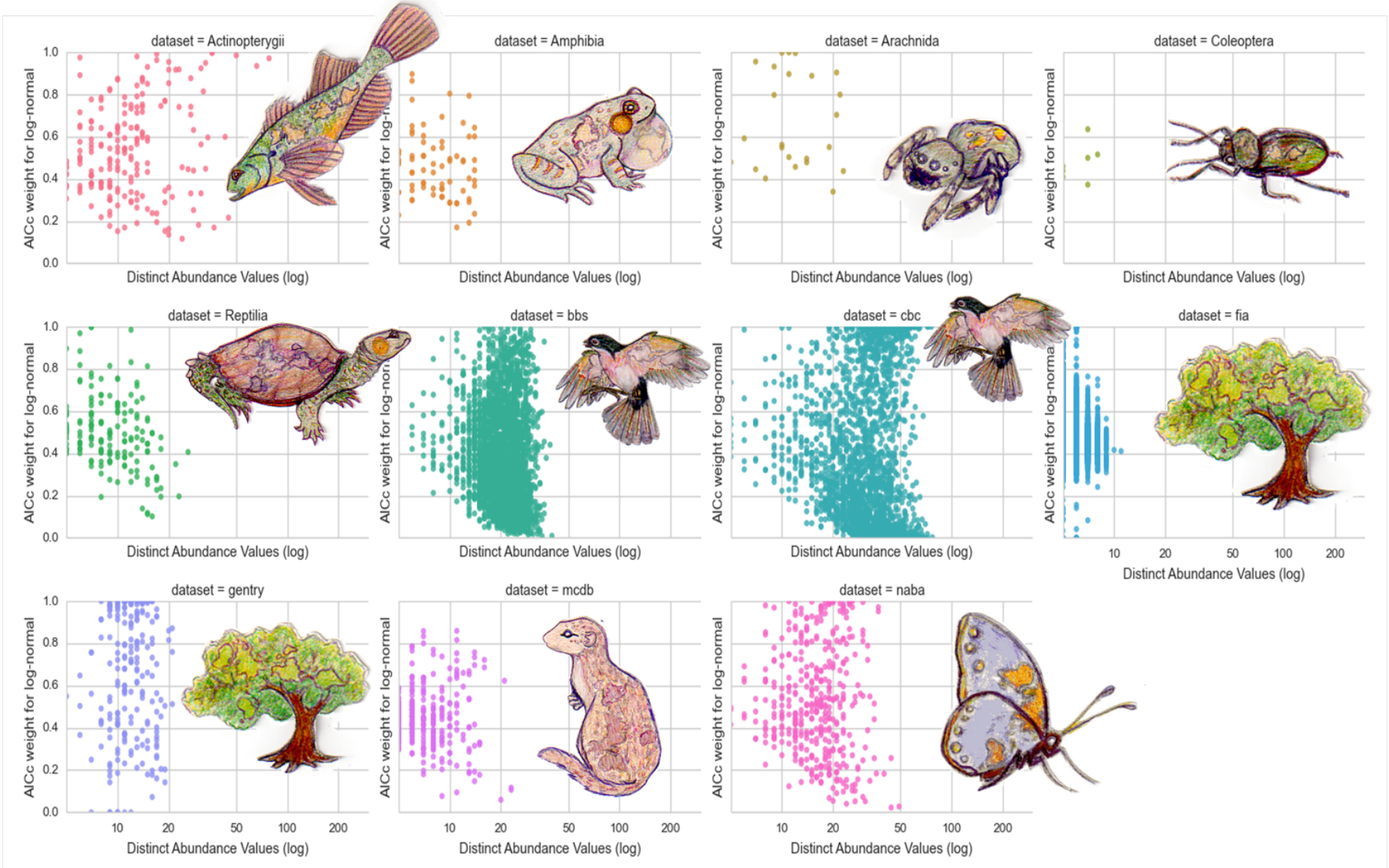


Connolly et al. 2014.

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.
    - May be biological mechanisms responsible for differences between marine and terrestrial communities.

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.

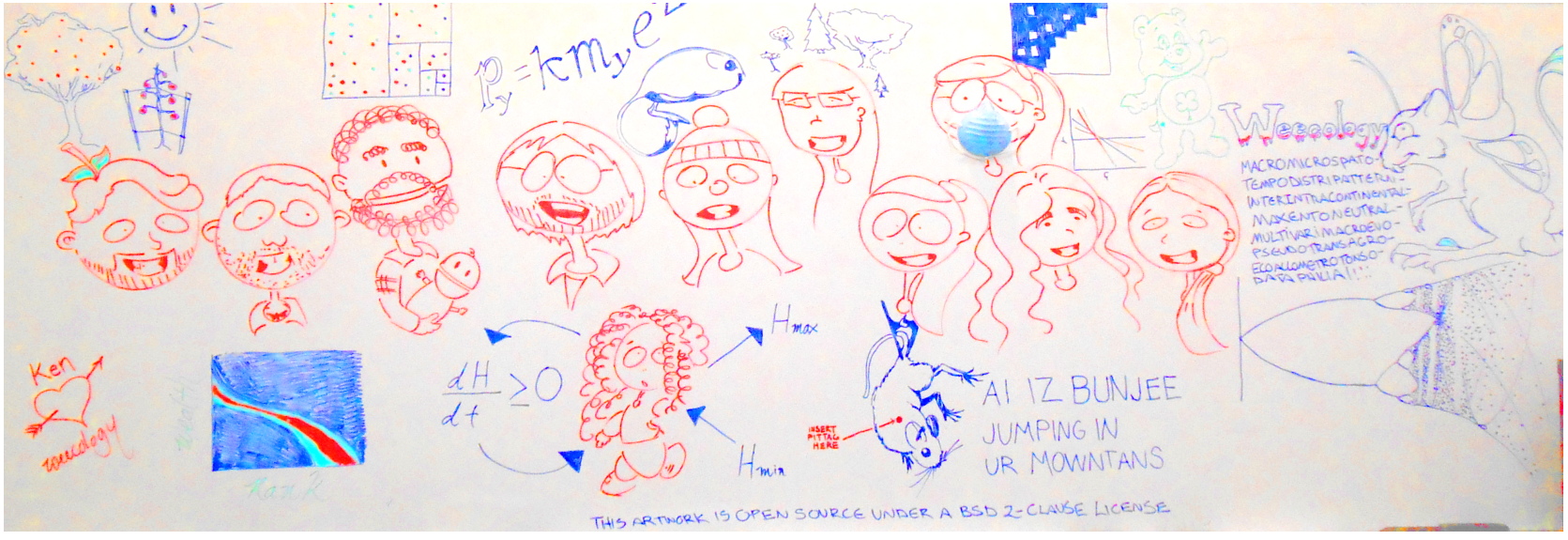
Acknowledgements

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.
- Accessibility information

Questions?