Linking Heterogeneous Data in Biodiversity Studies: *from field to database and the need for data carpentry*

Pamela S. Soltis
Florida Museum of Natural History
University of Florida

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

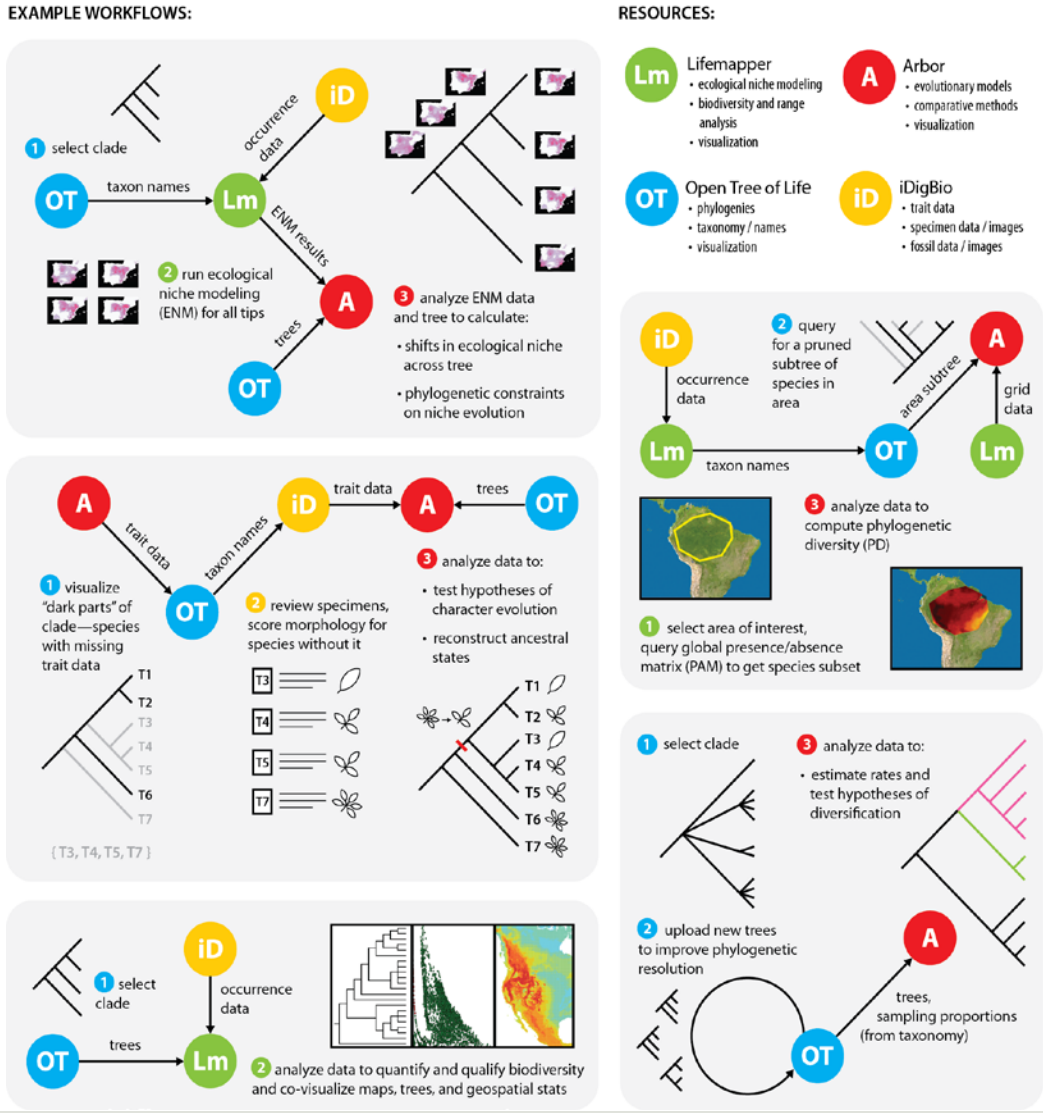
Synthetic Research and Complex Data Analysis

- Large data sets
- Heterogeneous data
- Multiple software packages
- Multi-step analytical workflows

- *Where do we get the training to do this?*

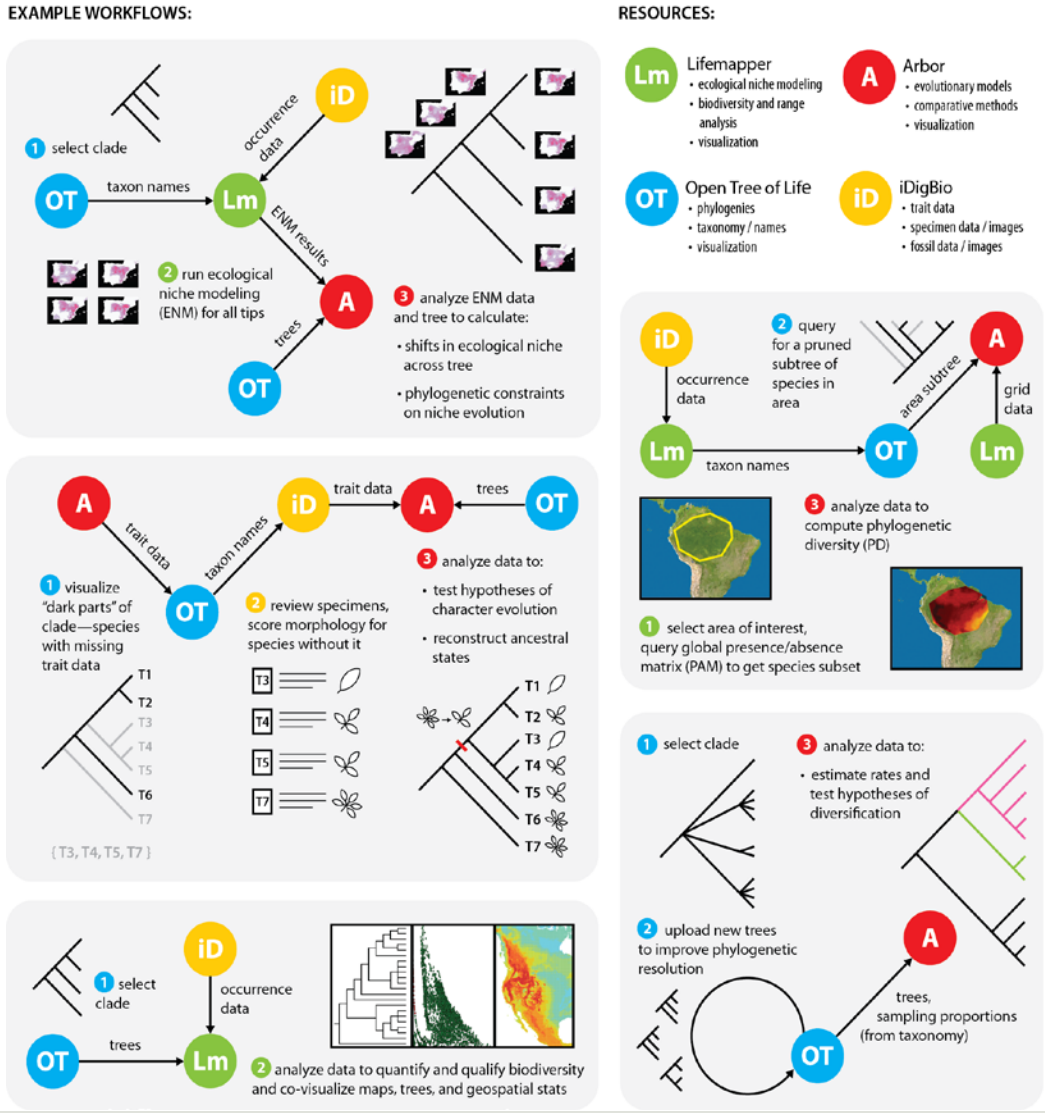
iDigBio
Integrated Digitized Biocollections

Connecting Trees, Specimens, Tools



Connecting Trees, Specimens, Tools

Connecting Trees, Specimens, Tools



Connecting Trees, Specimens, Tools

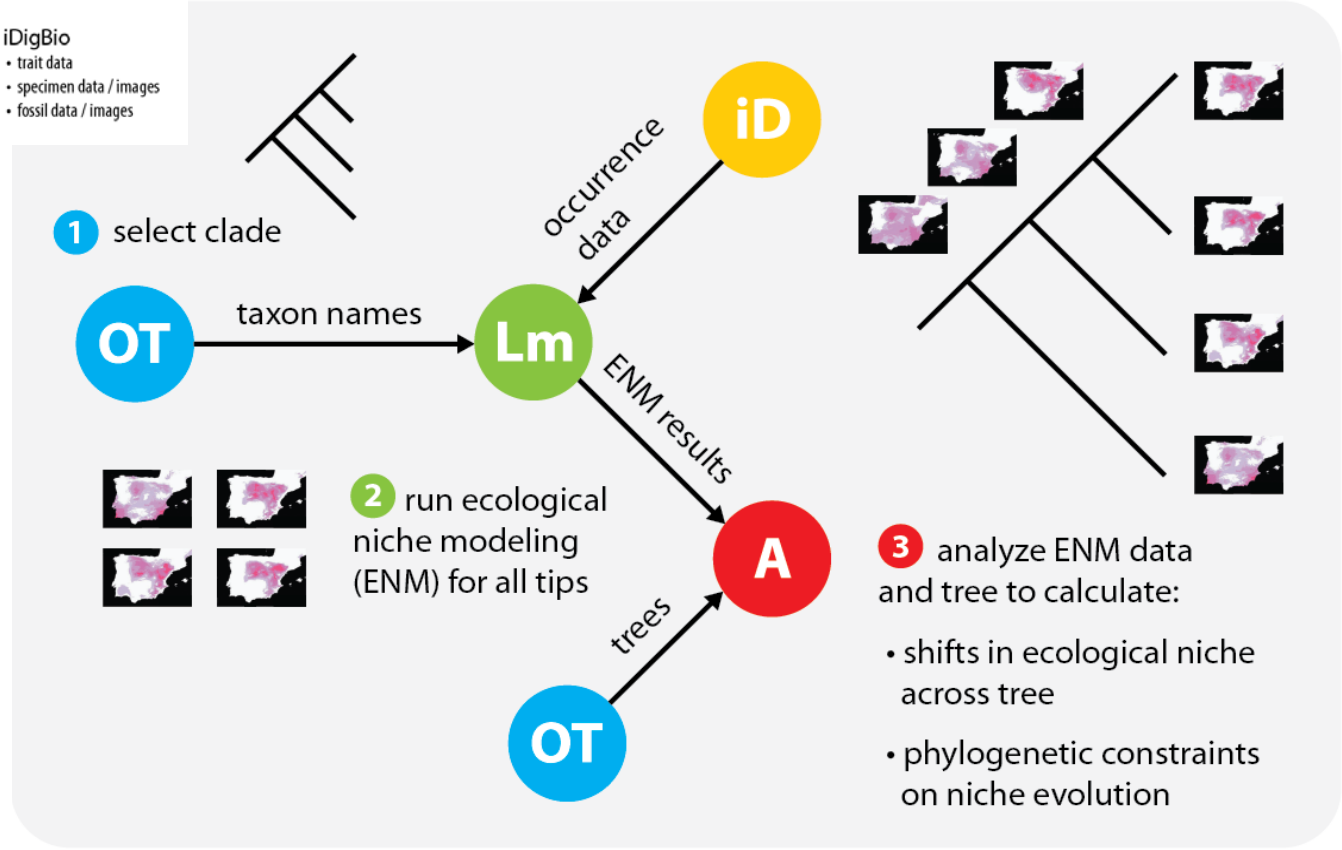
RESOURCES:

Lm Lifemapper
- ecological niche modeling
- biodiversity and range analysis
- visualization

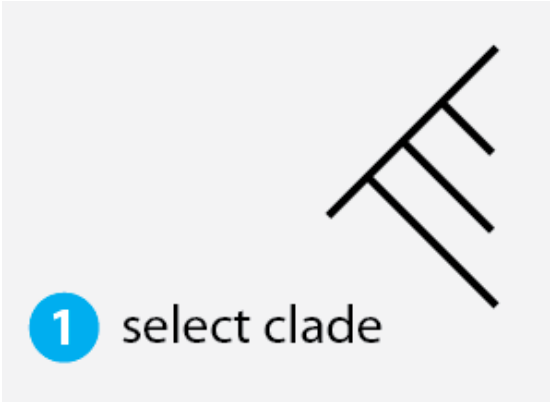
A Arbor
- evolutionary models
- comparative methods
- visualization

OT Open Tree of Life
- phylogenies
- taxonomy / names
- visualization

iD iDigBio
- trait data
- specimen data / images
- fossil data / images



Unpacking… Phylogenetic Trees

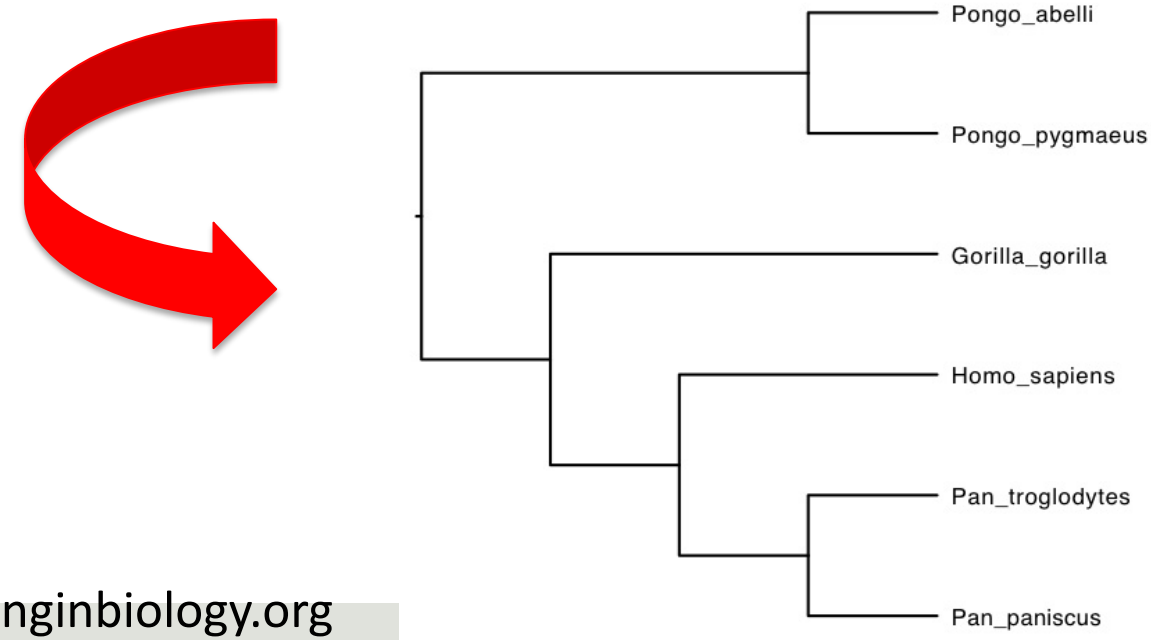


((cow:12, gnu:10)bigThings:3, (ant:23, bat:19)smallThings:5))

Unpacking… DNA Sequences

Using DNA Sequences to Build Trees

Characters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pongo abelli	A	T	G	A	C	C	T	C	A	A	C	A	C	G	T	A	A	A	T	C
Pongo pygmaeus	A	T	G	A	C	C	C	C	A	A	T	A	C	G	C	A	A	A	A	C
Gorilla gorilla	A	T	G	A	C	C	C	C	T	A	T	A	C	G	C	A	A	A	A	C
Homo sapiens	A	T	G	A	C	C	C	C	A	A	T	A	C	G	C	A	A	A	A	T
Pan troglodytes	A	T	G	A	C	C	C	C	A	A	C	A	C	G	C	A	A	A	A	T
Pan paniscus	A	T	G	A	C	C	C	C	A	A	C	A	C	G	C	A	A	A	A	T



Connecting Trees, Specimens, Tools

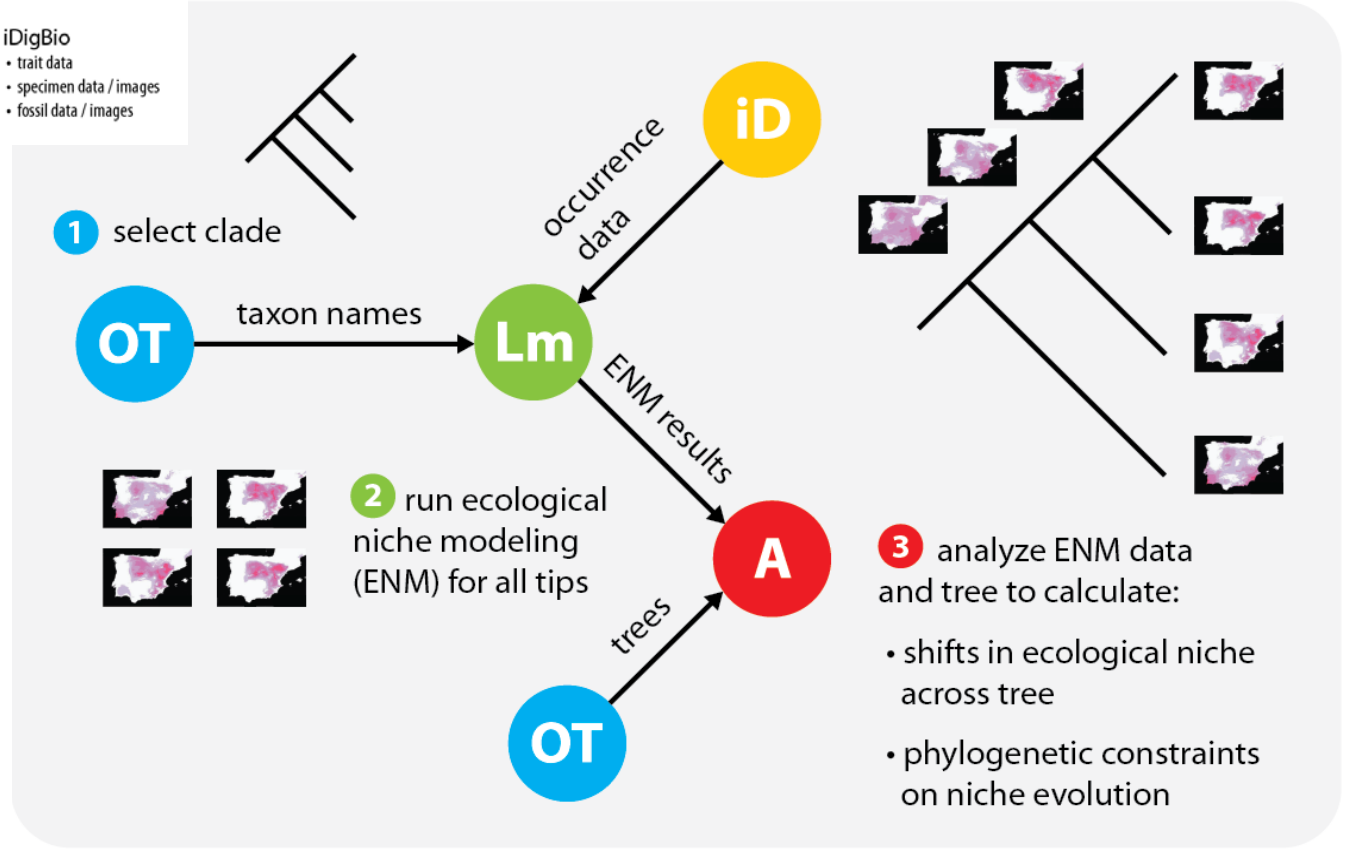
RESOURCES:

Lm Lifemapper
- ecological niche modeling
- biodiversity and range analysis
- visualization

A Arbor
- evolutionary models
- comparative methods
- visualization

OT Open Tree of Life
- phylogenies
- taxonomy / names
- visualization

iD iDigBio
- trait data
- specimen data / images
- fossil data / images

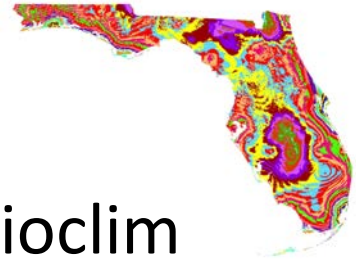


Unpacking…Ecological Niche Modeling: locations

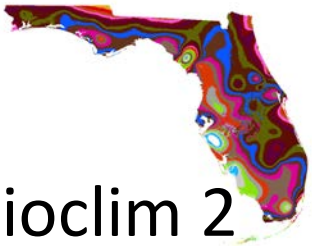
GEOLocate 29.65, -82.32

number,dwc:preparations,dwc:identificationVerificationStatus,idigbio:subfamily,idigbio:preparationCount,fcc:pickedBy,dwc:eventRemarks,dwc:VerbatimEventDate,dwc:associatedReferences,idigbio:endangeredStatus,dwc:locationAccordingTo,dwc:georeferenceSources,dwc:associatedSequences,dwc:formation,dwc:higherClassification,dwc:catalogNumber,dwc:verbatimSRS,dwc:higherGeography,dwc:individualCount,dwc:decimalLongitude,dwc:datasetName,dwc:month,dwc:georeferencedBy,dwc:eventTime,dwc:identificationQualifier,idigbio:

Unpacking…Ecological Niche Modeling: Environmental Layers

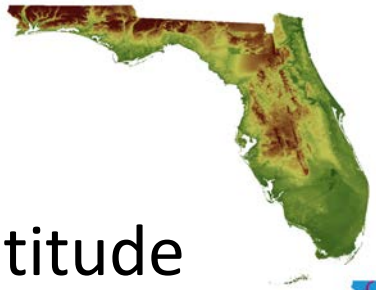


Bioclim 1



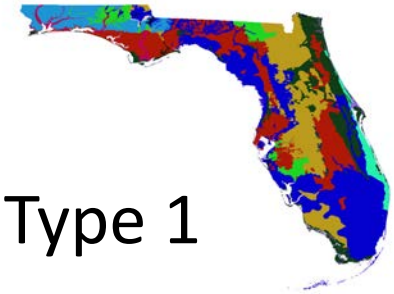
Bioclim 2

Bioclim 3

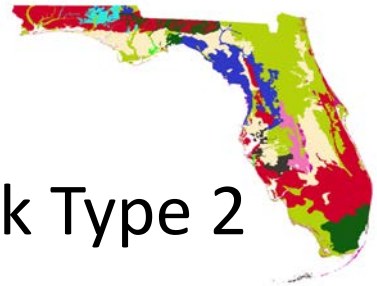


Altitude

Land Cover



Rock Type 1



Rock Type 2

C. Germain-Aubrey et al.

Connecting Trees, Specimens, Tools

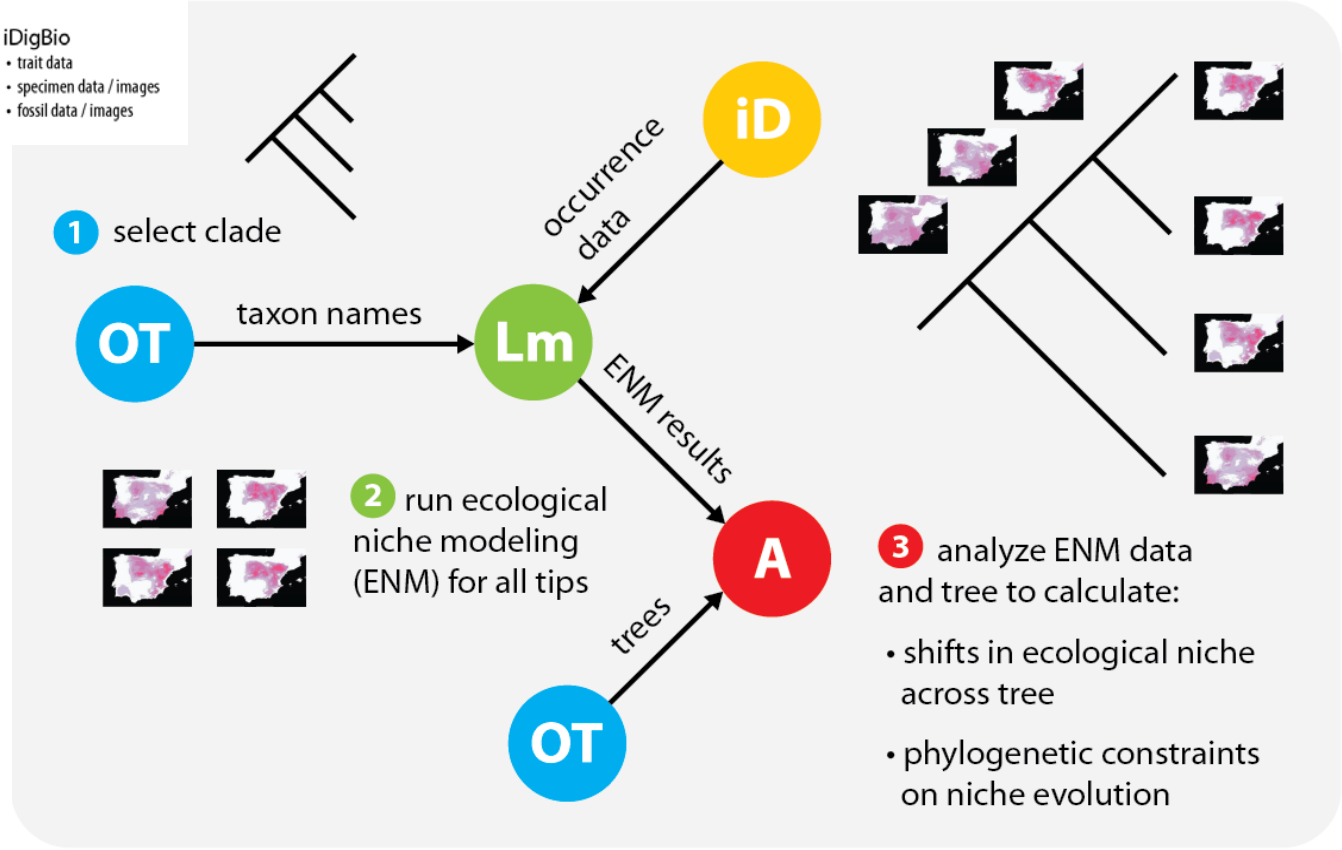
RESOURCES:

Lm Lifemapper
- ecological niche modeling
- biodiversity and range analysis
- visualization

A Arbor
- evolutionary models
- comparative methods
- visualization

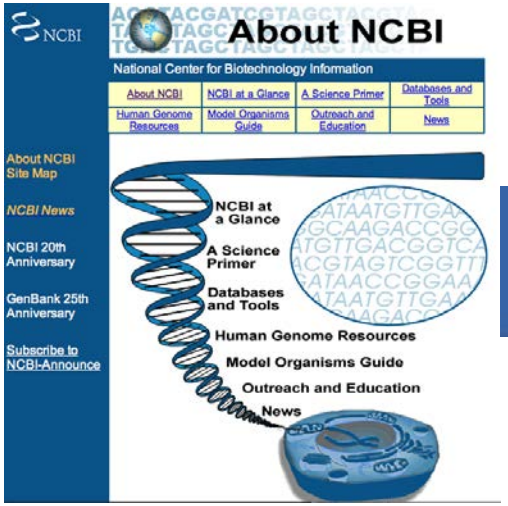
OT Open Tree of Life
- phylogenies
- taxonomy / names
- visualization

iD iDigBio
- trait data
- specimen data / images
- fossil data / images



Data Sources

- GenBank
- TreeBASE
- Dryad
- iDigBio
- Herbarium consortia
- Worldclim



TreeBASE
A Database of Phylogenetic Knowledge

iDigBio
Integrated Digitized Biocollections

DRYAD

WorldClim - Global Climate Data
Free climate data for ecological modeling and GIS

SERNEC
Southeast Regional Network of Expertise and Collections

Data Carpentry: skills…

- Necessary for synthetic research
- Important for employment
- Facilitate data sharing, reproducibility
- Prevent/reduce errors
- Prevent loss of data, wasted effort & funds





Data Carpentry: skills…

- Necessary for synthetic research
- Important for employment
- Facilitate data sharing, reproducibility
- Prevent/reduce errors
- Prevent loss of data, wasted effort & funds

Starting with Data Collection:

- Use/re-use
- Standards
- Management

Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, Aleksandra Slavkovic

Published: April 24, 2014 • DOI: 10.1371/journal.pcbi.1003542 • Featured in PLOS Collections

- Love your data, and help others love your data, too
- Share your data online, with a permanent identifier
- Conduct science with a particular level of reuse in mind
- Publish workflow as context
- Link your data to your publications as often as possible
- Publish your code (even the small bits)
- State how you want to get credit
- Foster and use data repositories
- Reward colleagues who share their data properly
- Be a booster for data science

Thanks and good luck!

www.idigbio.org

psoltis@flmnh.ufl.edu

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.