Analyzing paleoecological data: Best practices and current resources

Jessica Blois, Scott Farley, with contributions from Simon Goring, Jack Williams, Eric C. Grimm, and Edward Davis

9 January 2017 | IBS 2017 | Tucson, AZ

Learning Objectives

Workshop participants will learn:

- Key elements of paleoecological data
- how to use **APIs** (Application Programming Interface) for Neotoma and other databases
- how to use the **neotoma R package** to write scripts to retrieve Neotoma data directly into R for further analysis

Agenda

- 1 - 1:30pm: Getting started
  - Introductions
  - Getting set up for the day
- 1:30 - 2:00pm: Overview of paleoecological data
- 2:00 - 2:30pm: Neotoma web explorer
- 2:30 - 300pm: Intro to API and JSON queries
- 3:00 - 4:00pm: R Neotoma

Getting started

Resources

- Following along today
  - Intro/Background slides can be viewed through this link: `https://speakerdeck.com/jessicablois/analyzing-paleo-data-intro-slides`
  - The primary workshop document can be viewed through this link: `https://cdn.rawgit.com/NeotomaDB/Workshops/master/SVP2016/R/Workshop_SVP2016_v0.1.html` # Need to update!!
- All slides and materials are on GitHub
  - The Neotoma Paleoecology Database `https://github.com/NeotomaDB`
  - The Workshops section `https://github.com/NeotomaDB/Workshops/tree/master/SVP2016`

Getting started

Installations

1. Download and install R (available at `https://cran.r-project.org/`)
2. Download and install RStudio Desktop (`https://www.rstudio.com/products/rstudio/download3/`)
3. Install the following R packages: **RJSONIO**, **RCurl**, **neotoma**
   - Open up R Studio, then type: install.packages('RJSONIO', 'RCurl')
   - Install the `neotoma` package:

```
install.packages('devtools')
devtools::install_github("ropensci/neotoma")
```

Paleoecological data and the Neotoma Paleoecology Database

- We will discuss different types of paleo data in the context of Neotoma `www.neotomadb.org`, a database that houses different data from the more recent periods of Earth history

What is Neotoma?

- Neotoma is a database *consortium*: `www.neotomadb.org`
- Organized around proxy types
  - historically, brought together several major databases or datasets (i.e. FAUNMAP, COHMAP)
  - moving towards direct entry of individual data files
- The different databases use a common, extendable platform to ensure compatibility across proxy types
- Each proxy community can develop own standards, e.g. for taxonomy

The Neotoma ecosystem

- Faunal data
  - FAUNMAP, FAUNMAP2 (in progress)
  - MIOMAP (with Tony Barnosky)
  - ANTIGUA (South America megafaunal dates, in progress, with Tony Barnoksy and Emily Lindsey)
  - MQMD (Mexico, in progress, with Joaquin Arroyo Cabrales)
- Pollen and plant macros
- Packrat midden database
- Ostracodes, Diatoms
- Isotopes (coming online soon)
- Use `neotoma::get_table("datasettype")` to see all 25 available types

Scope and strengths of Neotoma

- Pliocene to present, global
- Chronology tools and storage
- Multiproxy
- Active development community
- Curated database –> high quality data

Curated database

- Each proxy group has a set of data stewards, led by one or a few lead stewards
- Anyone can deposit data into Neotoma
- Before going 'live', data need to be validated by a data steward

Key elements of paleoecological data

- Fundamental message: need to know the **taphonomy** of your study system in order to accurately infer biogeographic patterns across space and time
  - Occurrence records and taxonomy
  - Orienting occurrences in space
  - Orienting occurrences in time
- Some key references
  - Behrensmeyer AK, Kidwell S, Gastaldo R. 2000. Taphonomy and paleobiology. Paleobiology. 26(sp4):103–47
  - Jackson ST. 2012. Representation of flora and vegetation in Quaternary fossil assemblages: known and unknown knowns and unknowns. Quaternary Science Reviews. 49:1–16

Occurrence records and taxonomy

- Each occurrence is associated with a **taxonomic identification**
- Paleo data are often, though not always, less precise taxonomically than present-day data
- Different proxy types are recorded in different ways
  - Vertebrate data: MNI/NISP/Presence
  - Pollen data: originally counts per total pollen count, usually converted to relative abundance

Spatial precision

- lat-long coordinates, similar to present-day data
- **Different proxy types**, different taxa, and different depositional environments **record different levels of spatial precision**
  - Plant macrofossils vs pollen
  - Cave deposits vs fluvial deposits

Example: Fossil pollen from lake sediments

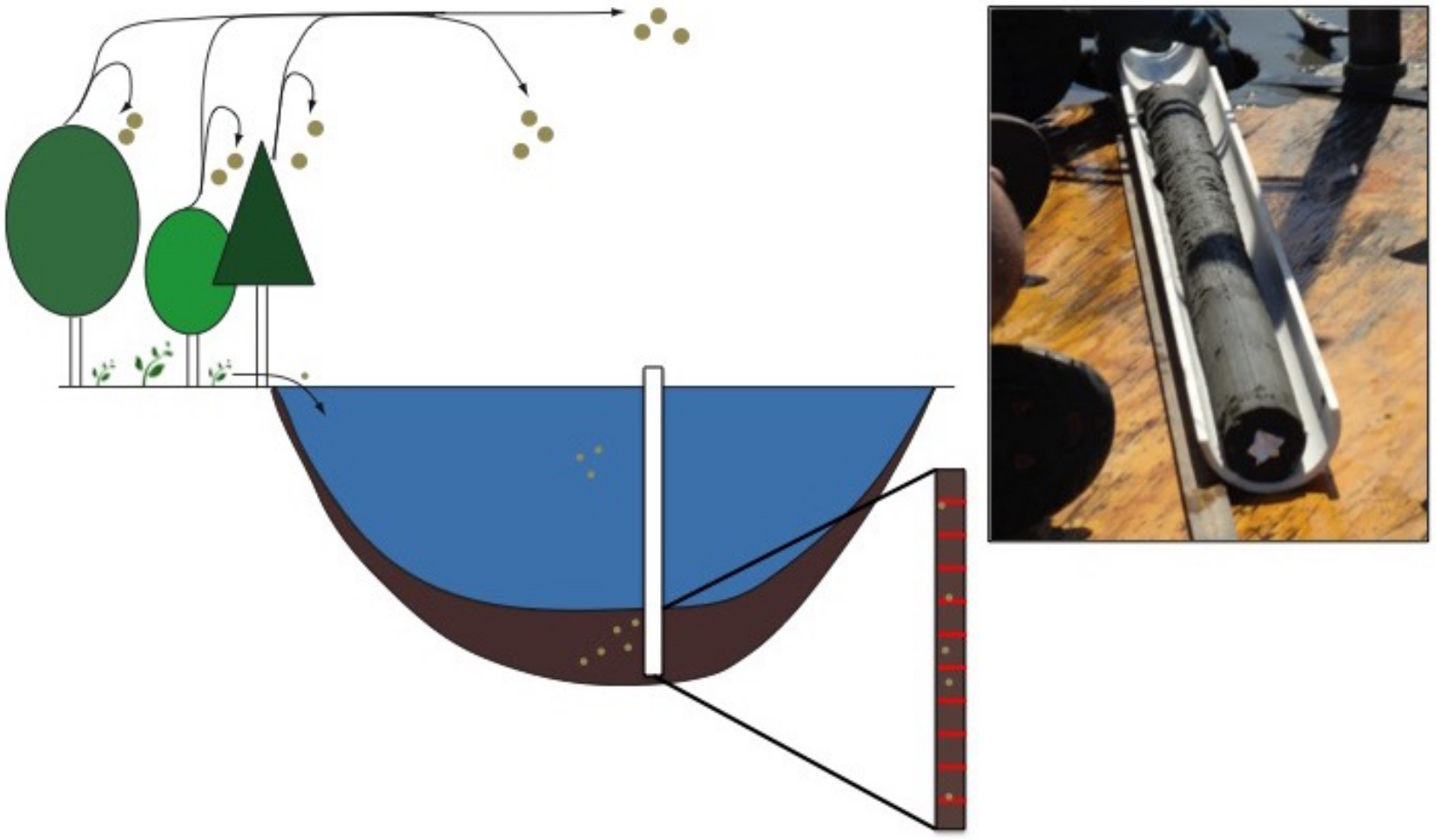

Figure 1:

Temporal precision

- Every occurrence is associated with an inference about the *time* the organism died
- There are different dating methods, each with their own assumptions and associated strengths and weaknesses
- **Different proxy types and depositional environments will be associated with different amounts of time-averaging**

Dataflow

Getting data out of Neotoma

- Focus of this workshop!
  - Neotoma Explorer
  - API
  - R
  - Database snapshots (housed at `www.neotomadb.org/snapshots`)

Key caveats*

- Errors or omissions in the data
- Updates to the data may not be stored in the database
  - new dates, changes in taxonomy, etc.
- "Garbage in, garbage out"
  - Many of the original databases like FAUNMAP and COHMAP convened a group of experts to assess various aspects of data quality. We need to keep applying that same lens to new datasets
  - Data without associated metadata can only go so far
    - *Note*: What is considered "good metadata" changes through time!
- When in doubt, record and store as much information as possible.

*for Neotoma and all other databases!

Major activities

- Adding new data (MIOMAP, ANTIGUA, MQMD)
- Integration with other databases: PBDB/Neotoma cross-database searches
- DOIs: assign datasets individual DOIs (will satisfy NSF data access guidelines)
- Embargos: tools for embargoing data.
  - will allow users to enter and validate data, get a DOI for paper submission, then data go 'live' when the paper is published
- Governance and Sustainability
  - Recently formed a leadership structure
  - Formalized bylaws and policy
- Education and outreach
  - Partnerships with SERC, Flyover Country, etc.

Questions and Discussion

- Thoughts, questions??
- Primary vertebrate contacts:
  - Jessica Blois, jblois@ucmerced.edu
  - Edward Davis, edavis@uoregon.edu
  - Russ Graham, rgraham@ems.psu.edu
- Primary pollen contacts:
  - Eric Grimm, eric.c.grimm@outlook.com
  - Jack Williams, jww@geography.wisc.edu
  - Simon Goring, simon.j.goring@gmail.com