We have all the data, now what?

The importance of planning for biodiversity data integration

Giovanni Rapacciuolo
Email giorapac@gmail.com
URL giorapacciuolo.com
Twitter @giorapac
Github @giorap

Defining Data Integration

Data integration involves:

1. Combining data residing at different sources

AND

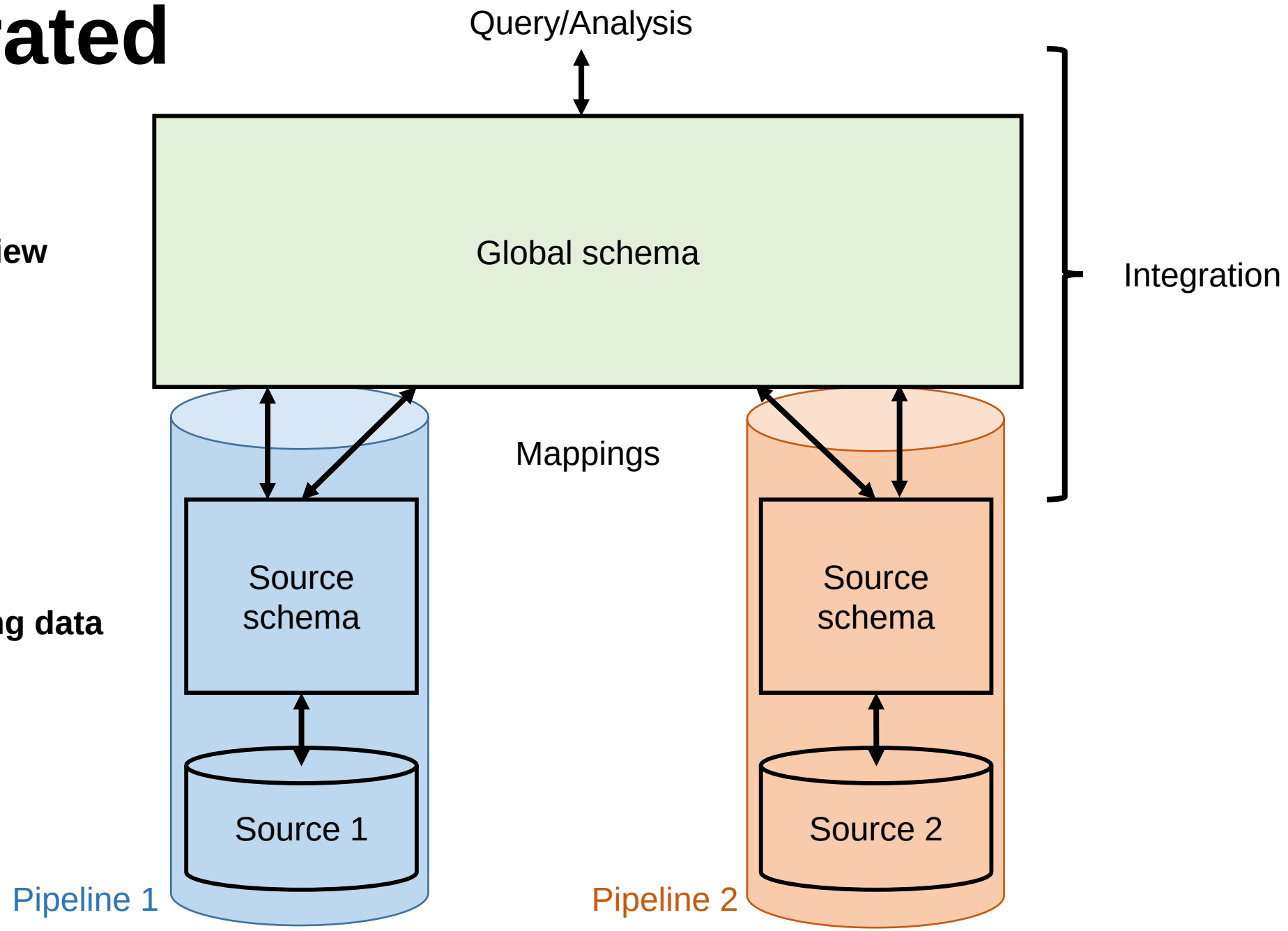
2. Providing users with a **unified view** of these data

Lenzerini (2002)

An integrated analysis

2. Unified view

1. Combining data

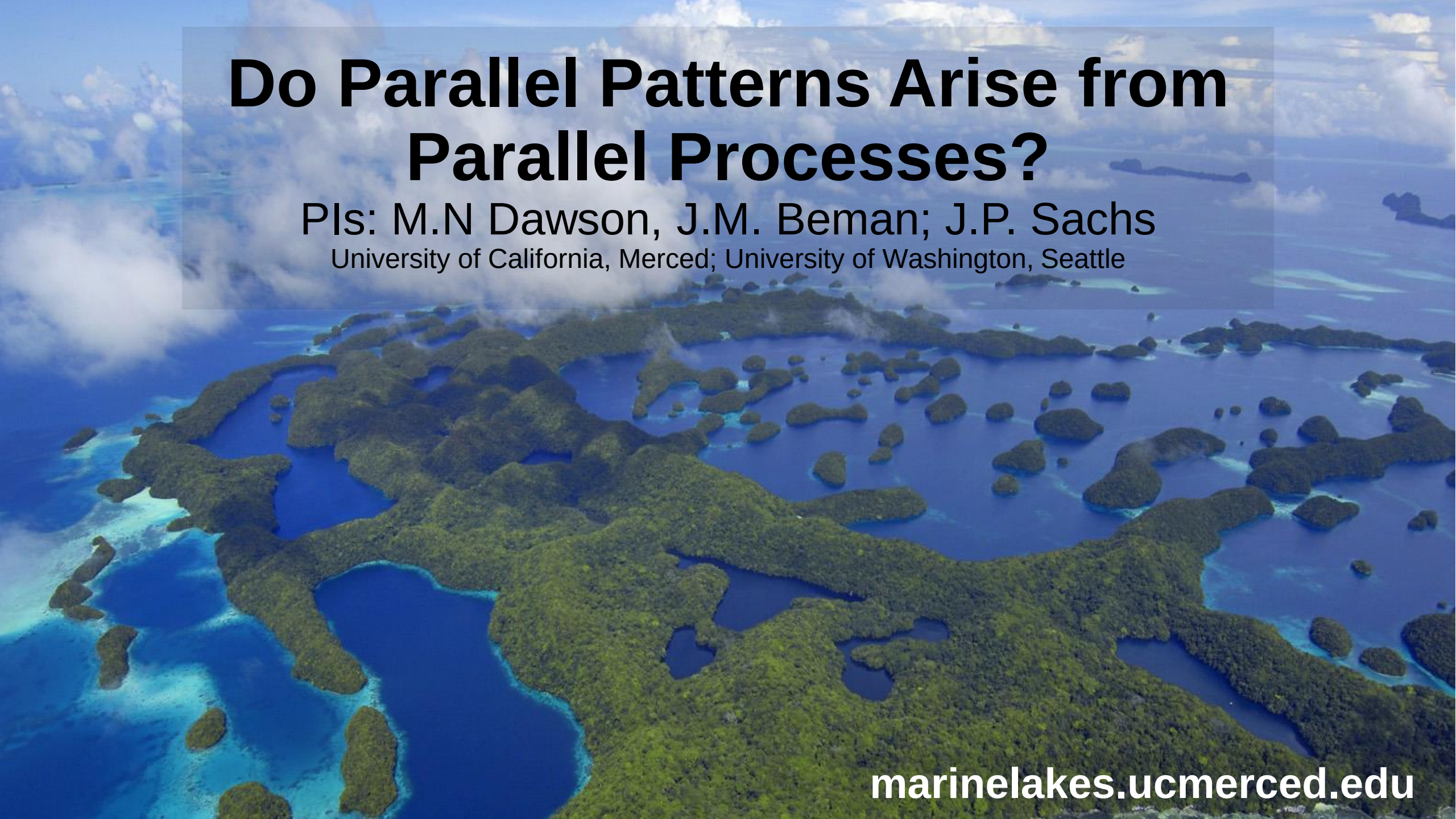


We need Data Integration Plans to address 3 key points

1. **WHY INTEGRATION**: Do we need data integration?

2. **WHEN INTEGRATION**: When do we need integration?

3. **HOW INTEGRATION**: How do we implement integration?



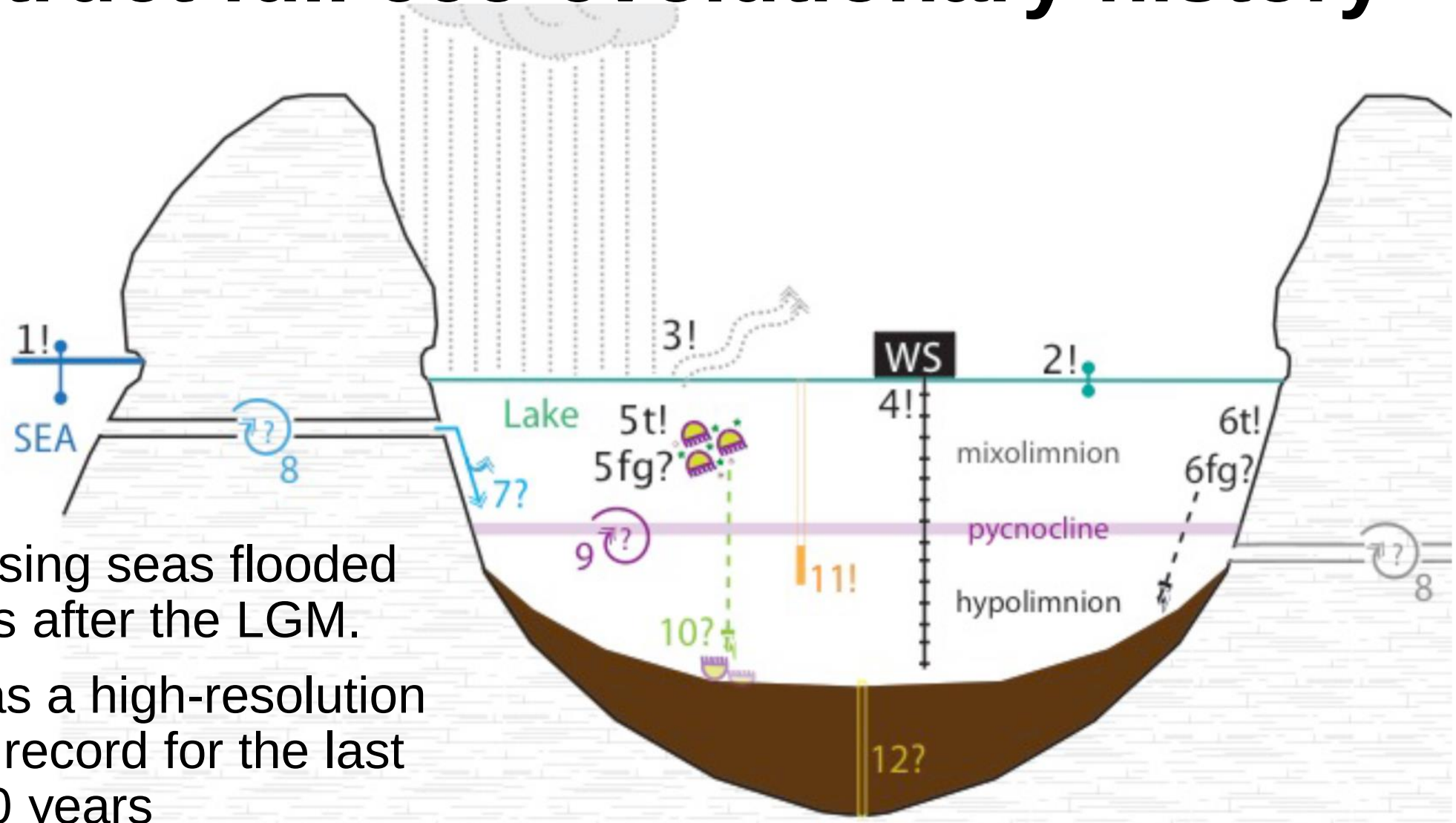
Do Parallel Patterns Arise from Parallel Processes?

PIs: M.N Dawson, J.M. Beman; J.P. Sachs

University of California, Merced; University of Washington, Seattle

marinelakes.ucmerced.edu

Marine lakes: a unique opportunity to reconstruct full eco-evolutionary history



- Formed as rising seas flooded inland valleys after the LGM.
- Each lake has a high-resolution sedimentary record for the last ~6000-15000 years

Three main data sources

1. Fish, invertebrate, plankton, microbial surveys
2. Sensor measurements of environmental variables
3. Physical and biological data from sediment cores

Why/when/how to integrate them?

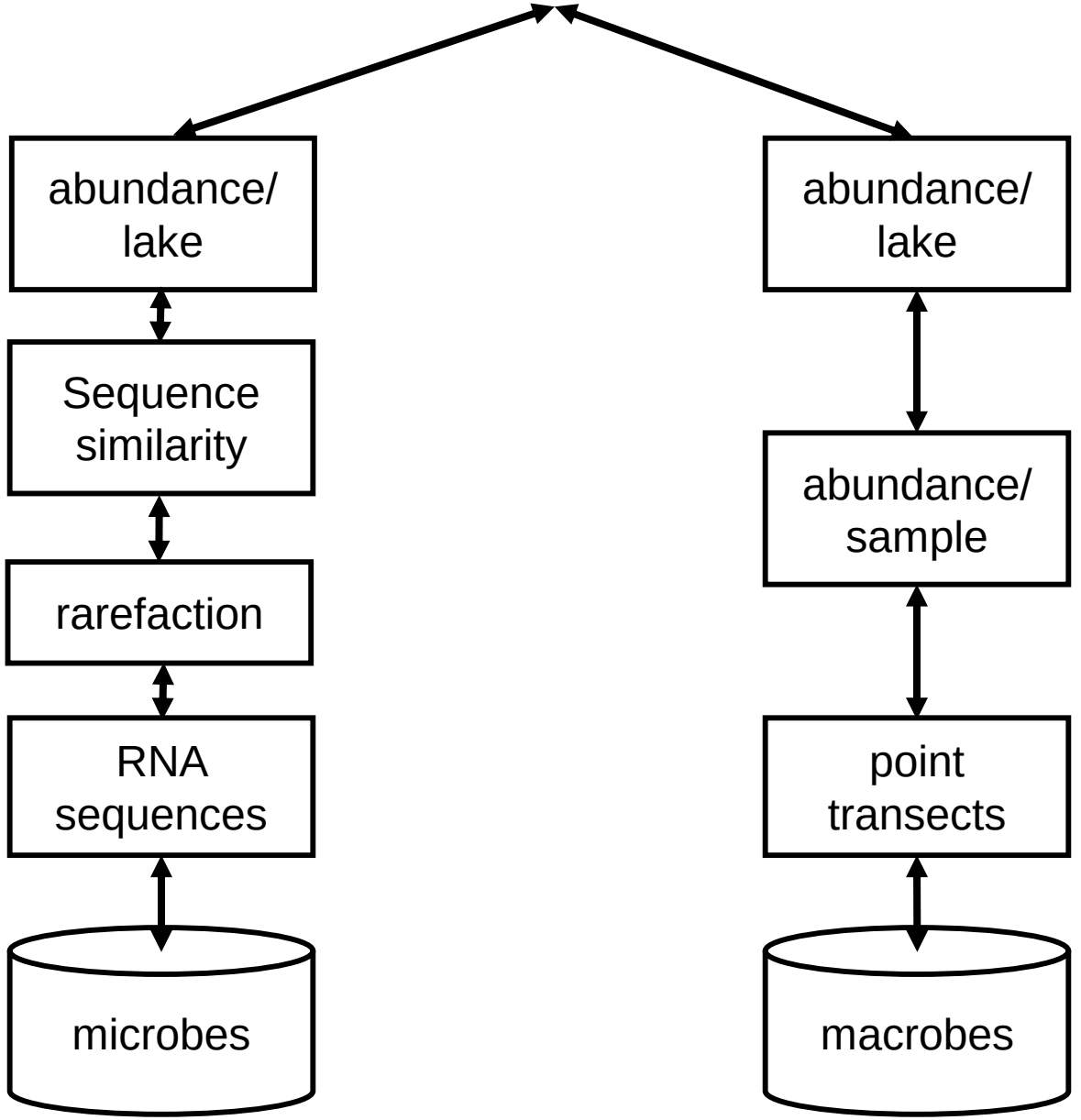
Example 1: comparing diversity patterns of microbes and macro-organisms

Query/Analysis lm(lake_microbial_richness ~ lake_macrobial_richness)

Source schemata

Data sources

Query/Analysis lm(lake_microbial_richness ~ lake_macrobial_richness)



Source schemata

Data sources

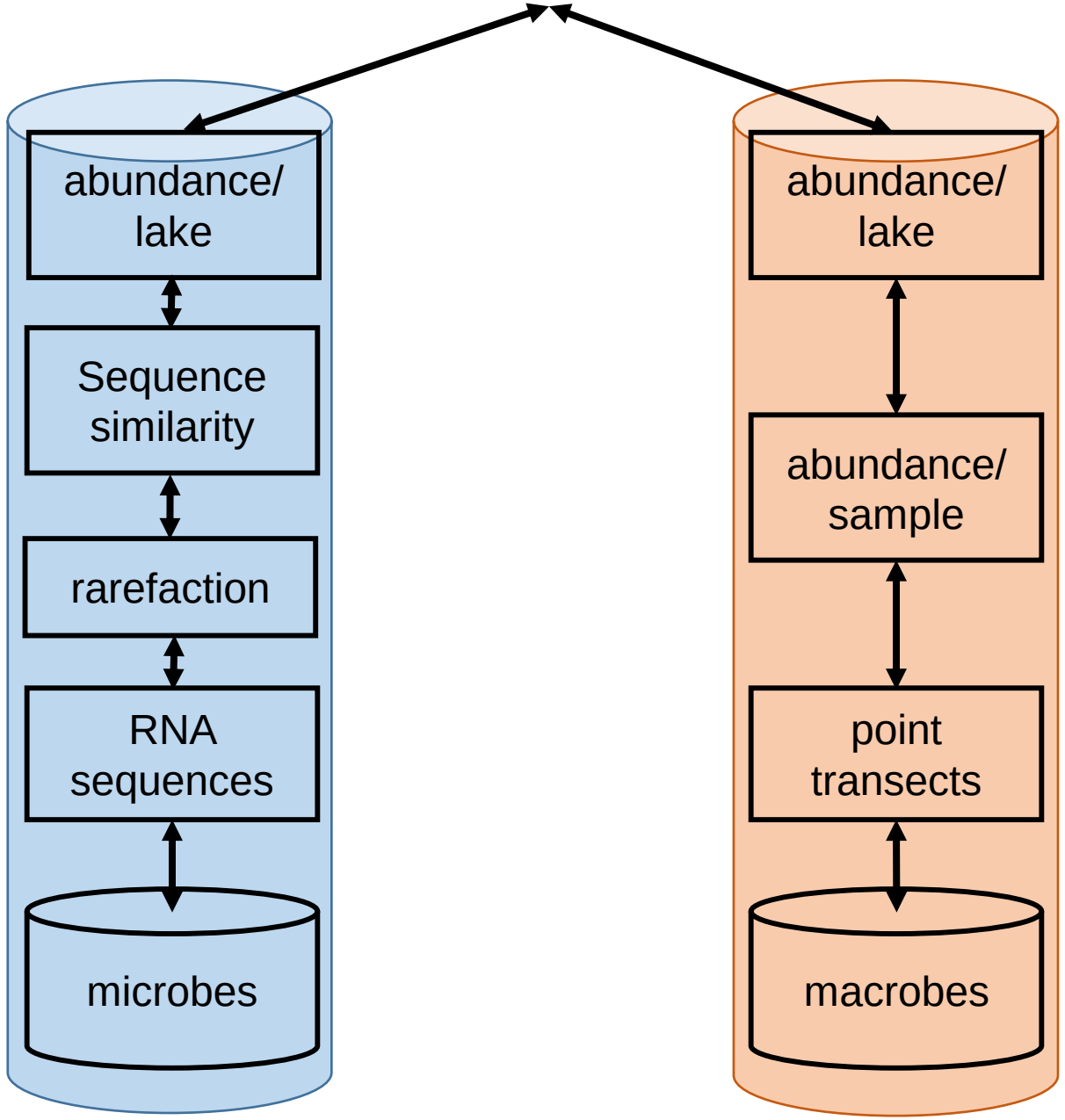
Query/Analysis

lm(lake_microbial_richness ~ lake_macrobial_richness)

Mappings

Source schemata

Data sources



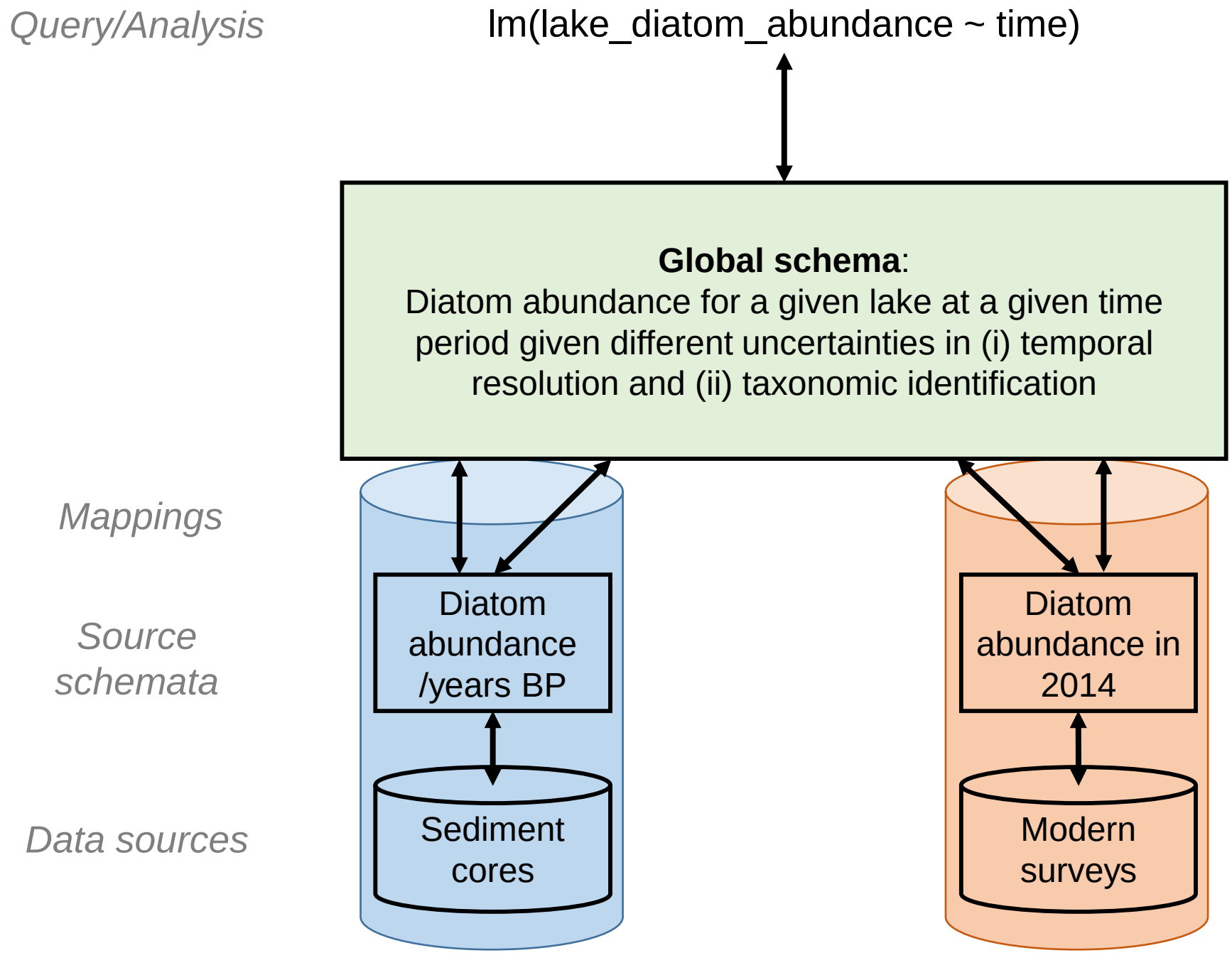
Example 2: diatom abundance changes from ~10,000 years ago until present

Query/Analysis

lm(lake_diatom_abundance ~ time)

Source schemata

Data sources


Query/Analysis
lm(lake_diatom_abundance ~ time)
Global schema:
Diatom abundance for a given lake at a given time period given different uncertainties in (i) temporal resolution and (ii) taxonomic identification
Mappings
Source schemata
Data sources
Diatom abundance /years BP
Sediment cores
Diatom abundance in 2014
Modern surveys

Data Integration Plans facilitate large-scale biodiversity data analyses

Data Integration Plans describe:

1. **All the data sources** involved (quality, consistency, uncertainty, etc.)
2. **Clearly defined goals**: the desired analytical outputs
3. **Why/When/How** to implement the process of integration to facilitate the desired outpus

Testing Data Integration Plans using simulation?

How about running **simulations** with dummy data to see IF we can integrate data once we have them?

We need Data Integration Plans to address 3 key points

1. **WHY INTEGRATION**

2. **WHEN INTEGRATION**

3. **HOW INTEGRATION**

Giovanni Rapacciuolo
Email giorapac@gmail.com
URL giorapacciuolo.com
Twitter @giorapac
Github @giorap