# SCIENTIFIC ANNOTATION

## USING GRAPHS TO FACILITATE INTERDISCIPLINARY SCIENCE

Data resources in the Earth Sciences range from kilobytes to terabytes, with a range of user communities, technical competence, and scientific uses. There is a need to increase usability and discoverability, to assign credit to data generators, developers and to provide people with the tools and knowledge they need to help address some of society's most pressing issues. This presentation showcases an approach that uses neo4j's graph database, open source APIs, the use of persistent identifiers such as ORCIDs and DOIs, and crowd sourcing to help connect disciplines within the Earth Sciences to undertake science for a new century.

Simon Goring, Assistant Scientist

University of Wisconsin - Madison

**goring.org** - @sjgoring

tweet #odscwest

# SCIENTIFIC ANNOTATION

## USING GRAPHS TO FACILITATE INTERDISCIPLINARY SCIENCE

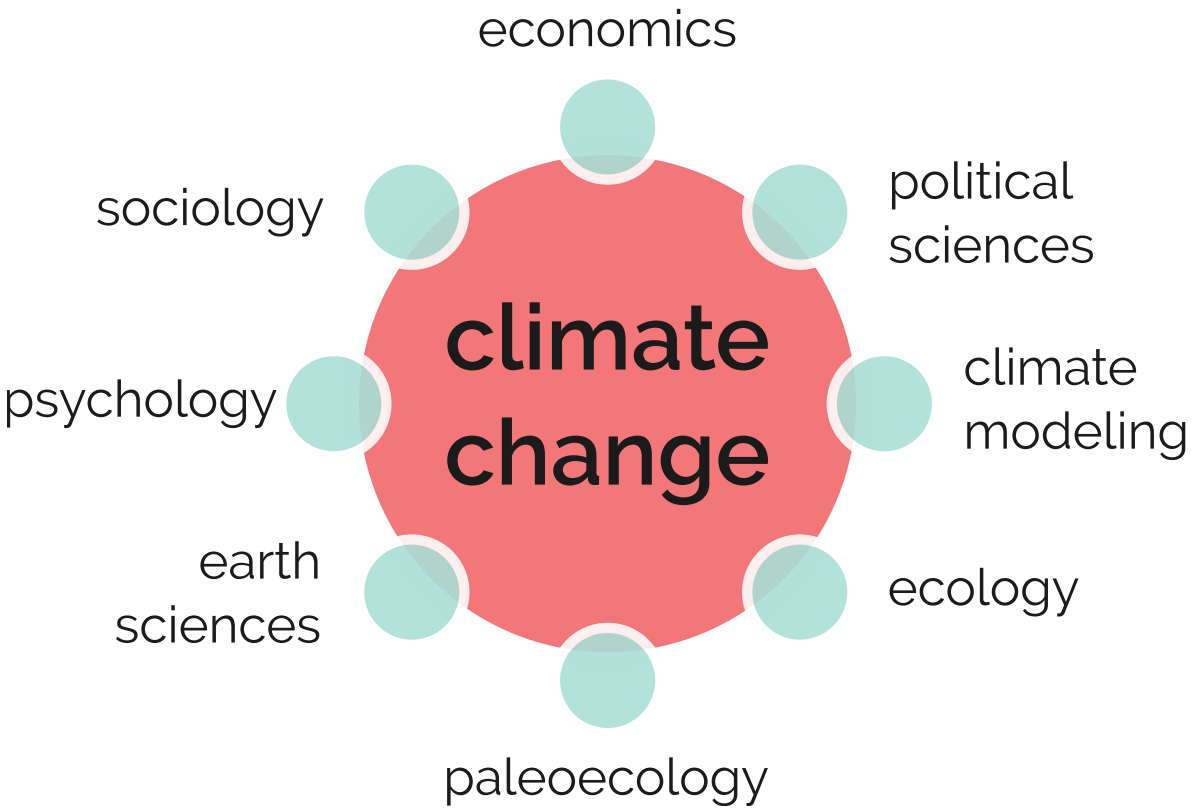
Simon Goring
Steve Richard, Kirsten Lehnert, Doug Fils, Nick McKay,
Steve Khuen, Anders Noren, Jack Williams
University of Wisconsin, Columbia University, Ocean Leadership,
Northern Arizona University, Concord College, University of Minnesota



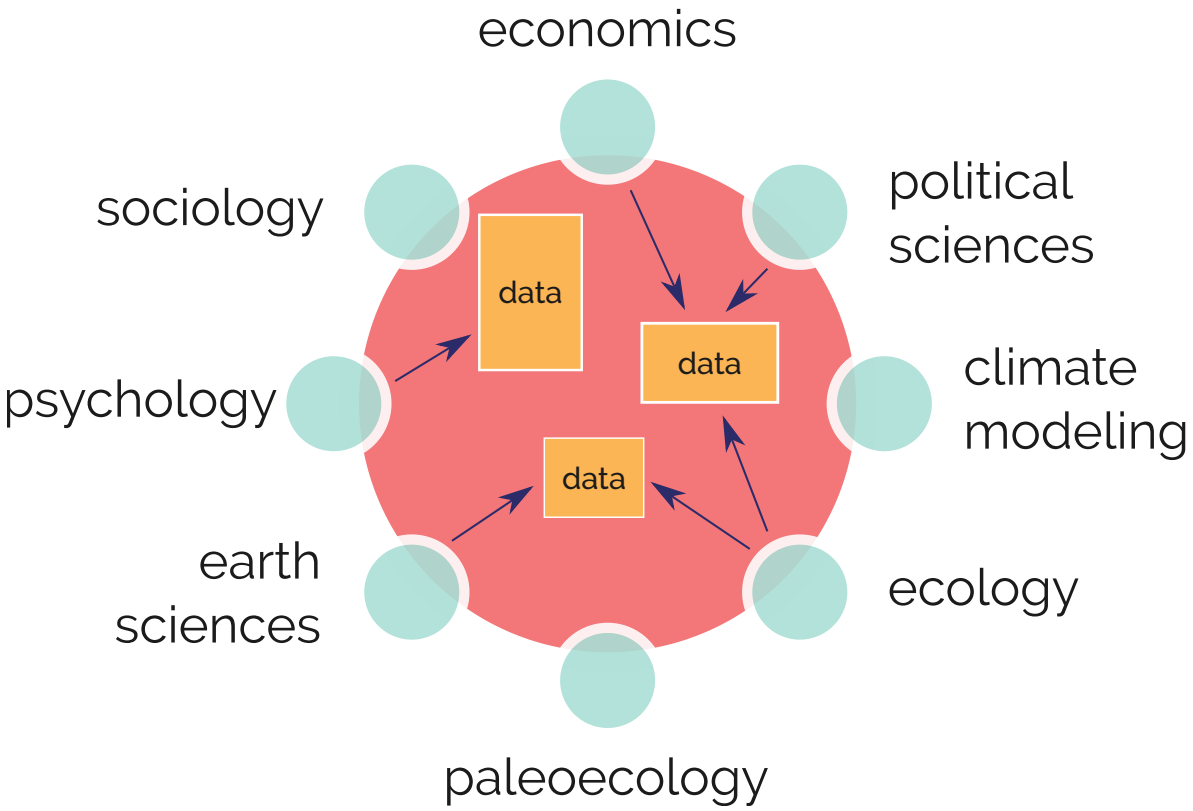
EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH




# WICKED PROBLEMS




# WICKED PROBLEMS




All recognize metadata's potential value, but when the rubber meets the road, an unfunded mandate to be altruistic . . . does not prove highly attractive. **Introduced in order to reduce data friction, metadata creates its own kind of friction**.

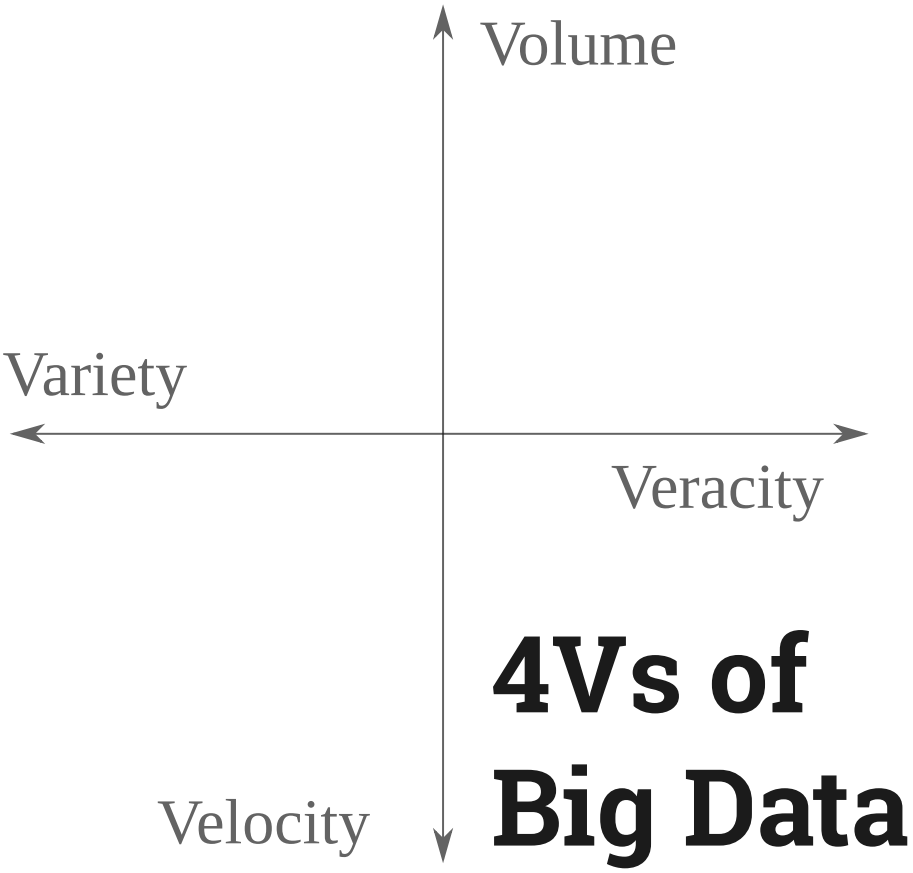
Edwards et al. 2011. Social Studies of Science

# BIG DATA IN THE LONG TAIL

## Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions

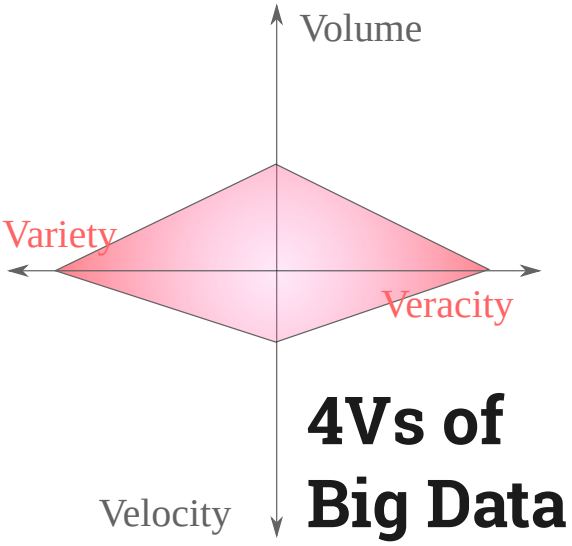
*Scott S. Farley, Andria Dawson, Simon J. Goring and John W. Williams*

*Bioscience* **68**:563–576. DOI: **10.1093/biosci/biy068**




**AXES OF BIG DATA**

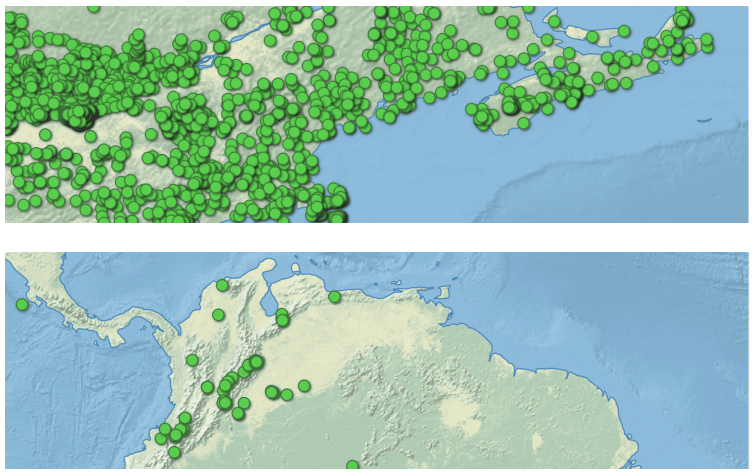
# COMMUNITY CURATED DATA RESOURCES

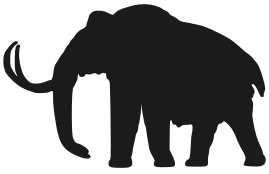




**Goring *et al.* (2018); Williams *et al* (2017); Farley *et al* (2017)**

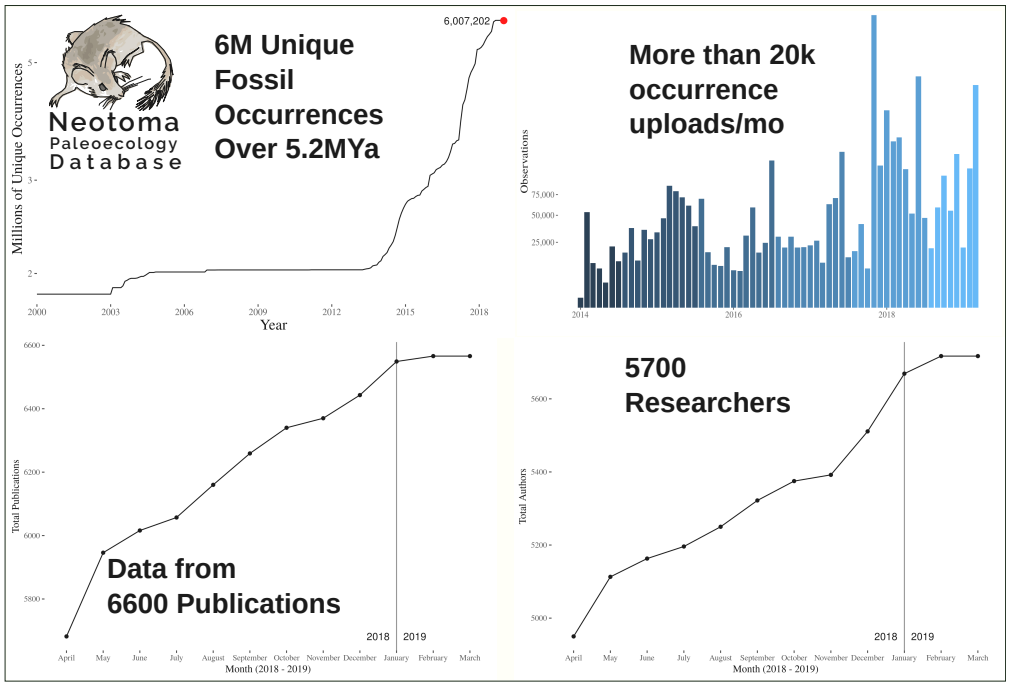
# NEOTOMA



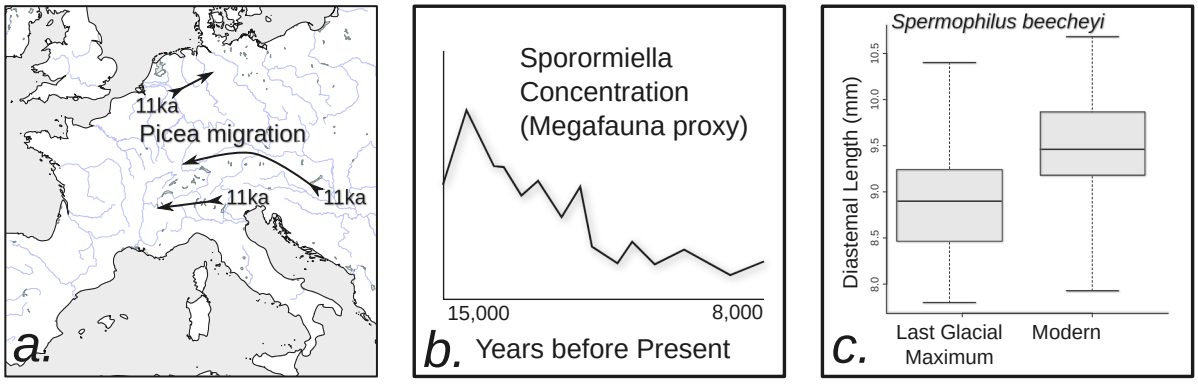


Neotoma Paleoecology Database - **neotoma.org**

# NEOTOMA DATABASE

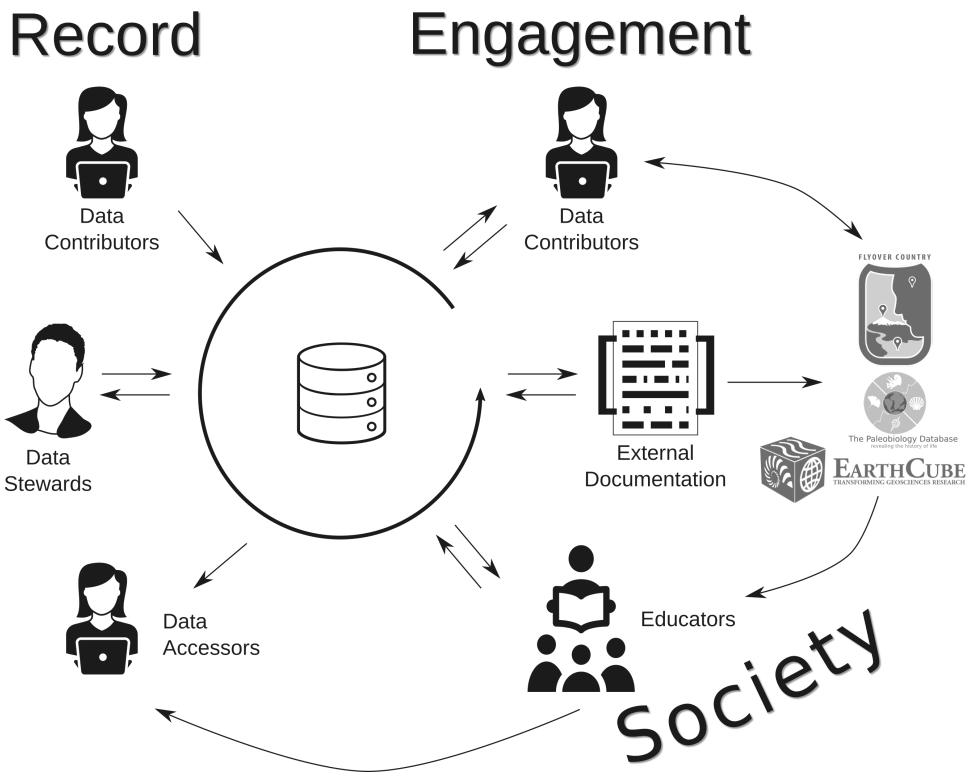


# NEOTOMA DATA

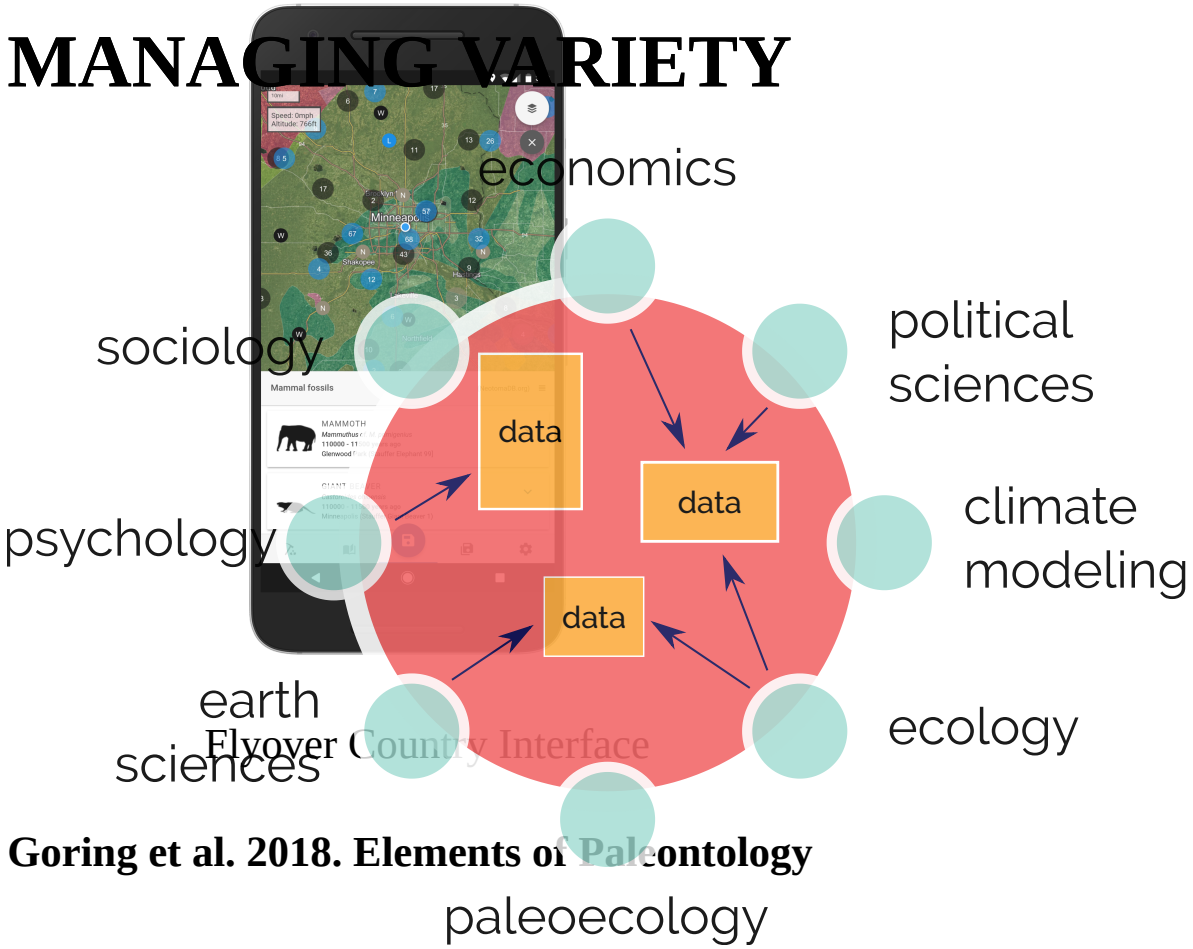



**Goring et al. 2018. Elements of Paleontology**

# COMMUNITY CURATED DATA RESOURCES



# MANAGING VARIETY

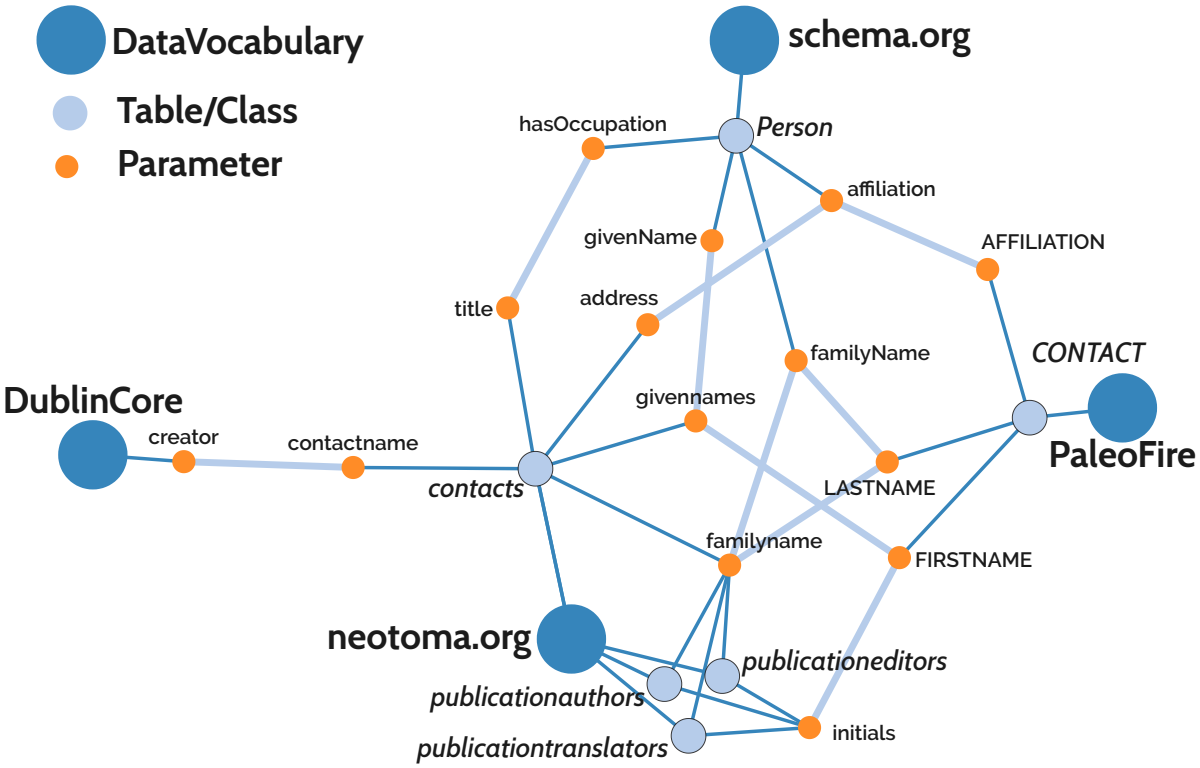



**Goring et al. 2018. Elements of Paleontology**

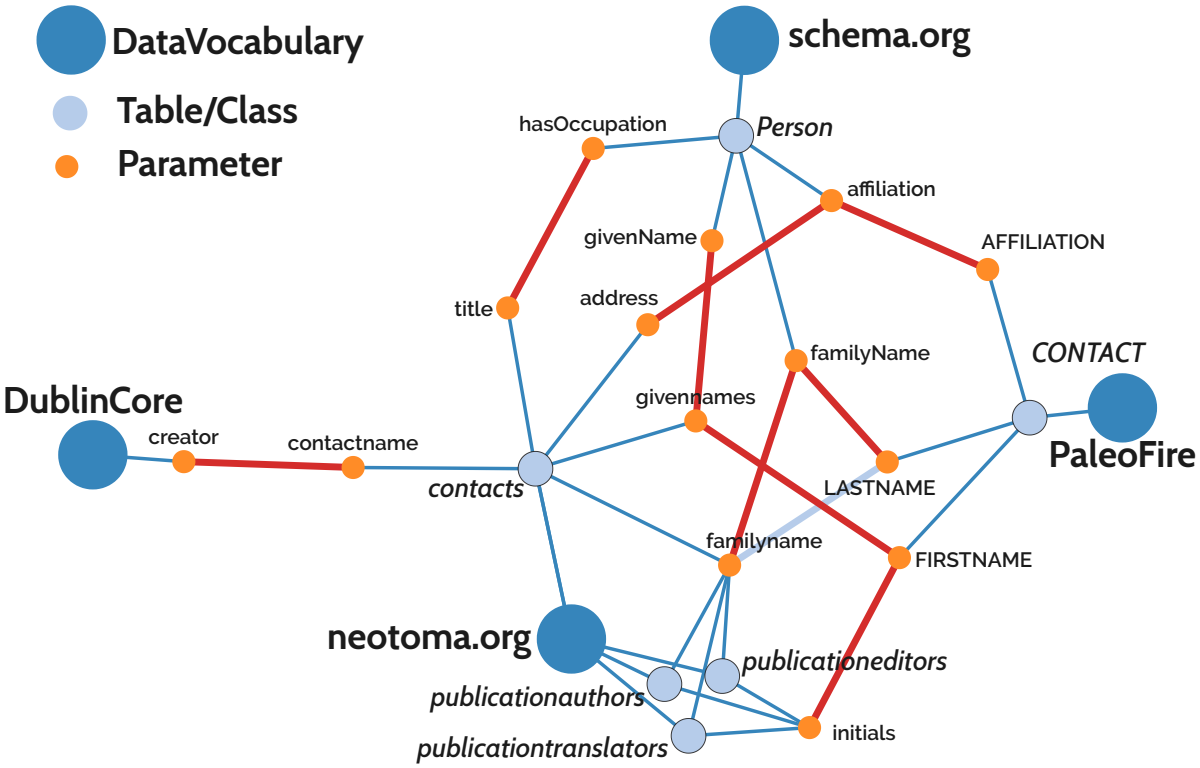
# MANAGING VARIETY

- Development and adoption of standards
    - **schema.org**, W3C
    - Khider et al. **2019**. Crowdsourced Reporting Standards for Paleoclimate Data.
- Cross-resource collaboration
    - EarthLife Consortium, Flyover Country
- Industry partnerships
    - Google Dataset Search/Project 418

# DATA ALIGNMENT

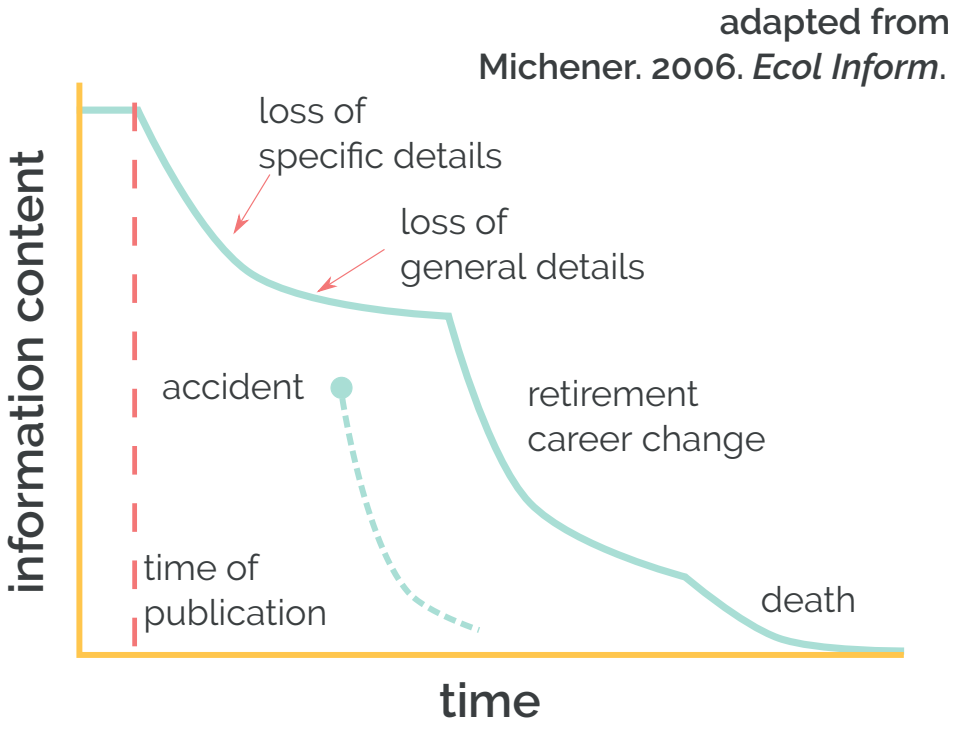


# DATA ALIGNMENT

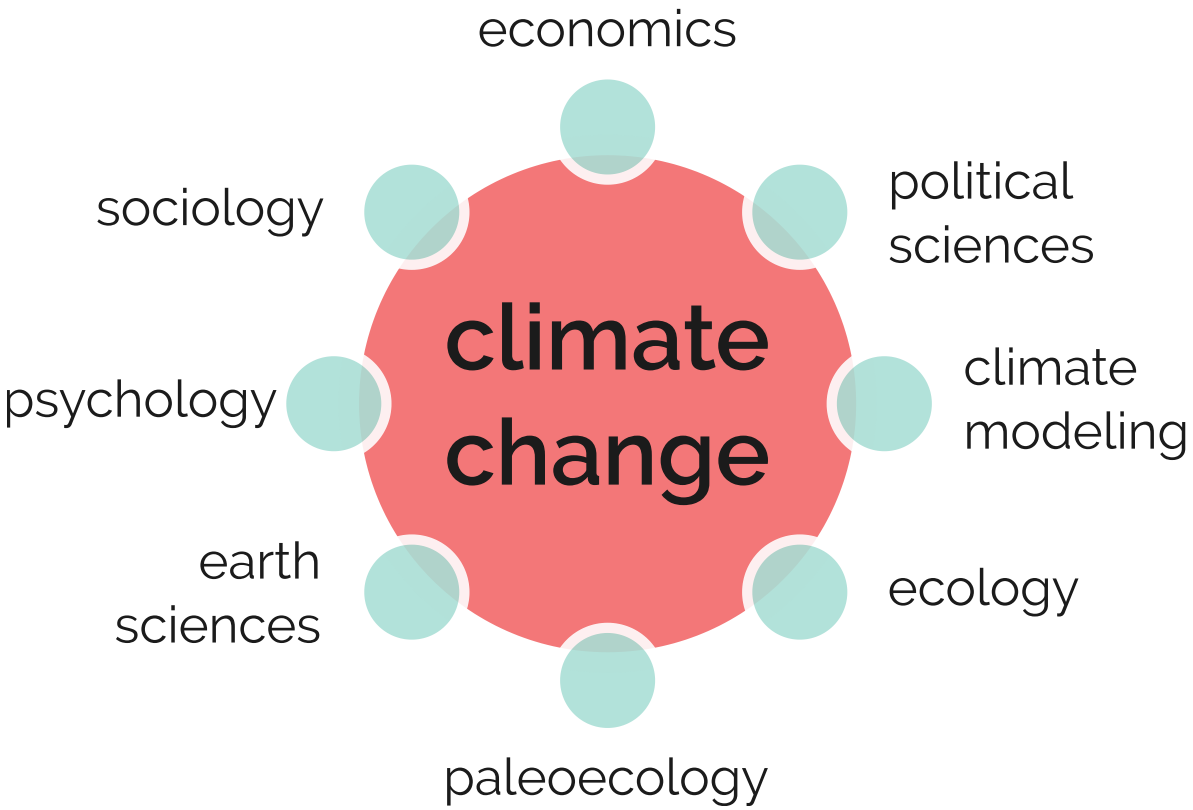


# DATA ALIGNMENT

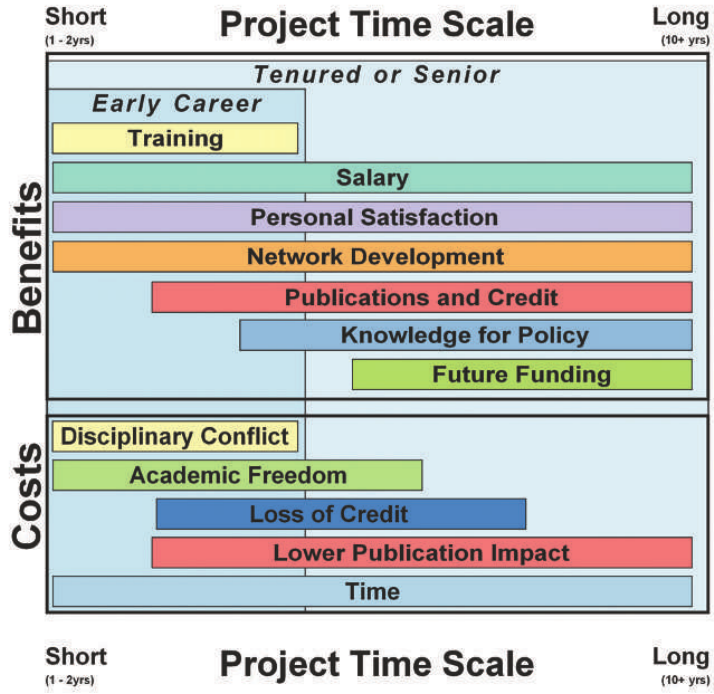
# MANAGING VERACITY



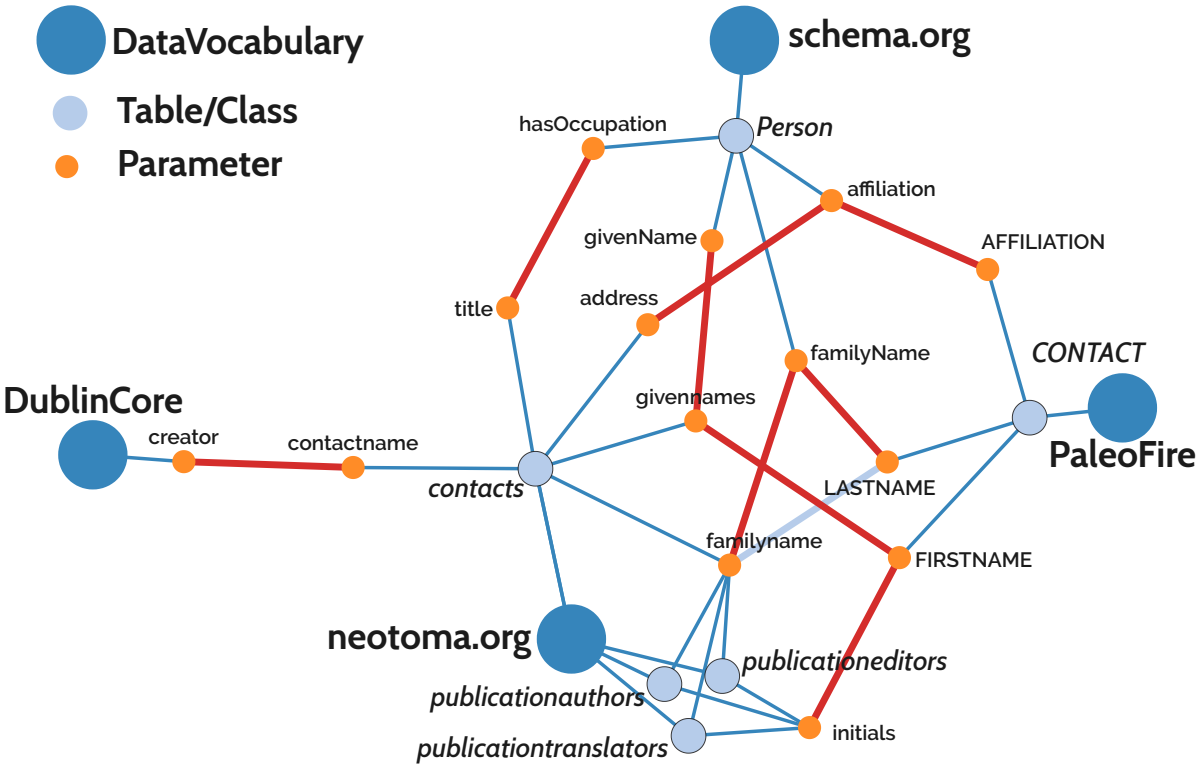
# WICKED PROJECTS




# INTERDISCIPLINARY COSTS



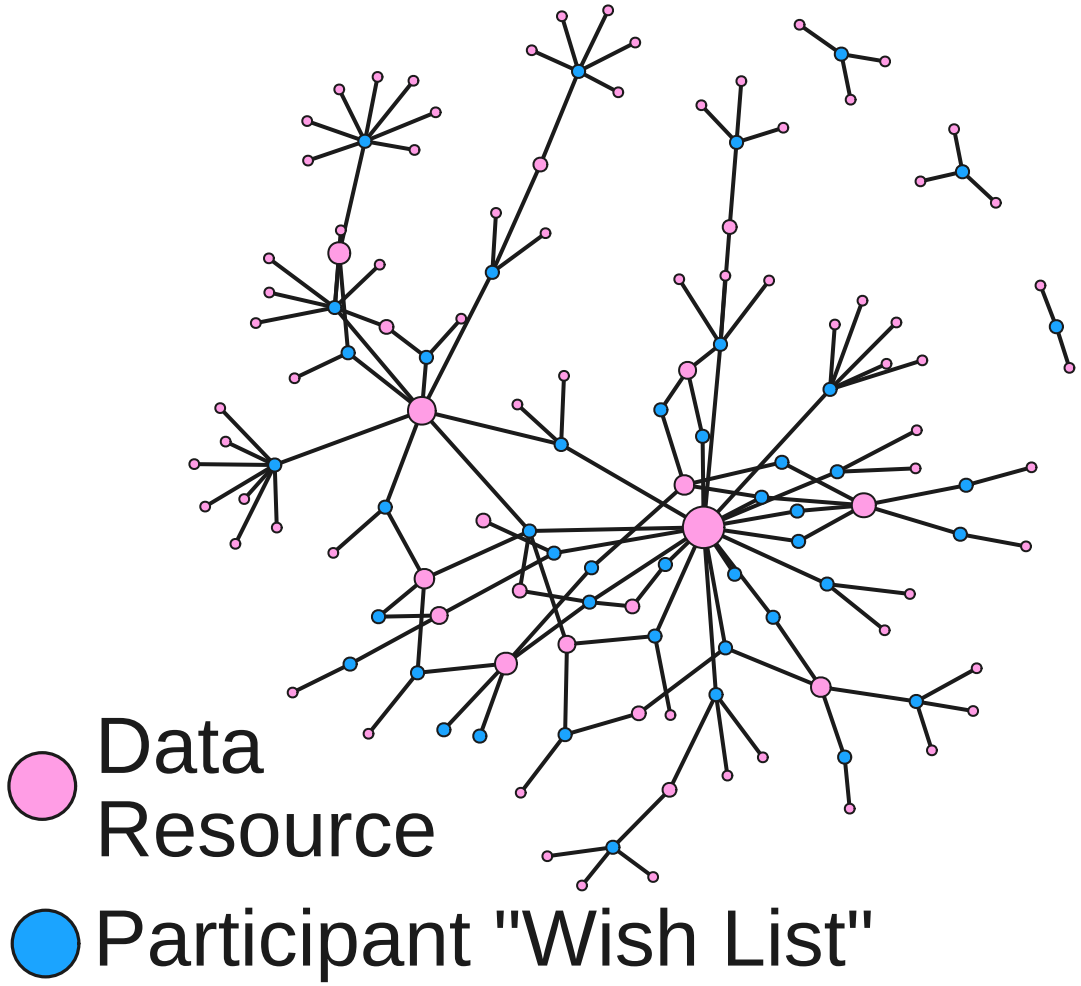
# DATA ALIGNMENT



# ANNOTATION GRAPHS

# EARTHCUBE AND THROUGHPUT

# CONNECTING RESOURCES



# ANNOTATION MODEL

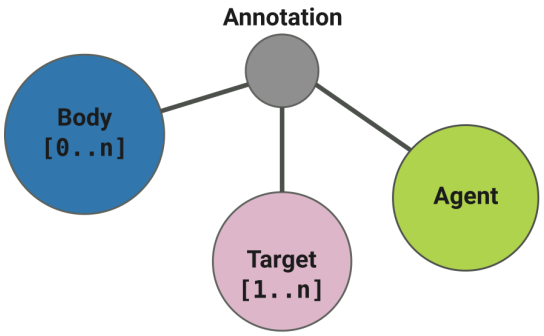
An Annotation is a rooted, directed graph representing a relationship between resources.

Two primary types of resource participate in an annotation, **Bodies** and **Targets**.

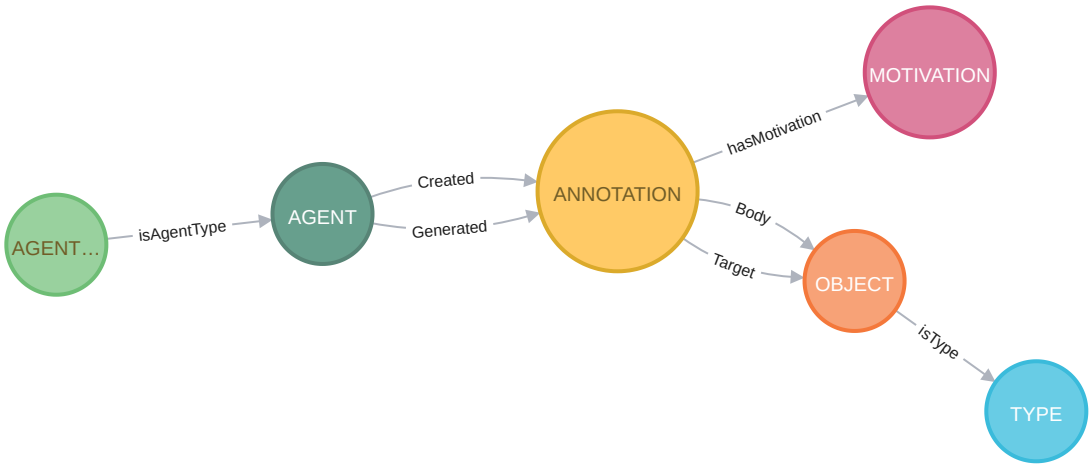
Body resources are related to, and typically "about", Target resources. Annotations, Bodies and Targets may have their own properties and relationships.

The intent behind the creation of an Annotation or the inclusion of a particular Body or Target is an important property and represented by a Motivation resource.

w3c Web Annotation Data Model



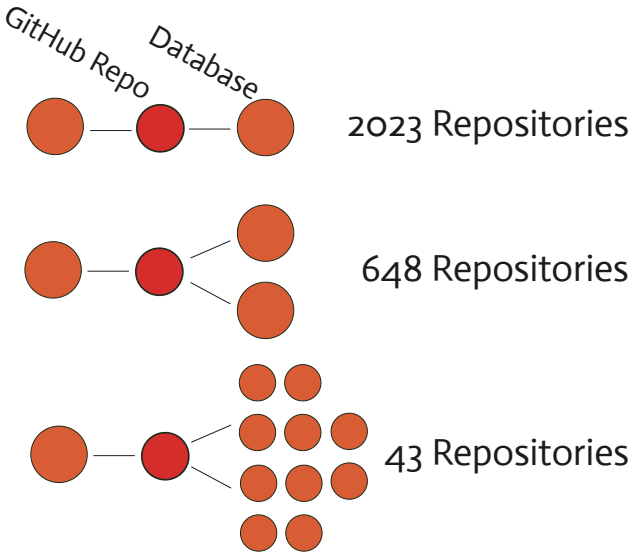
# ANNOTATION MODEL



## GITHUB REPOSITORIES

- 2301 Data Resources Catalogued
- 24,000 GitHub Repositories Linked

# CONNECTIVITY



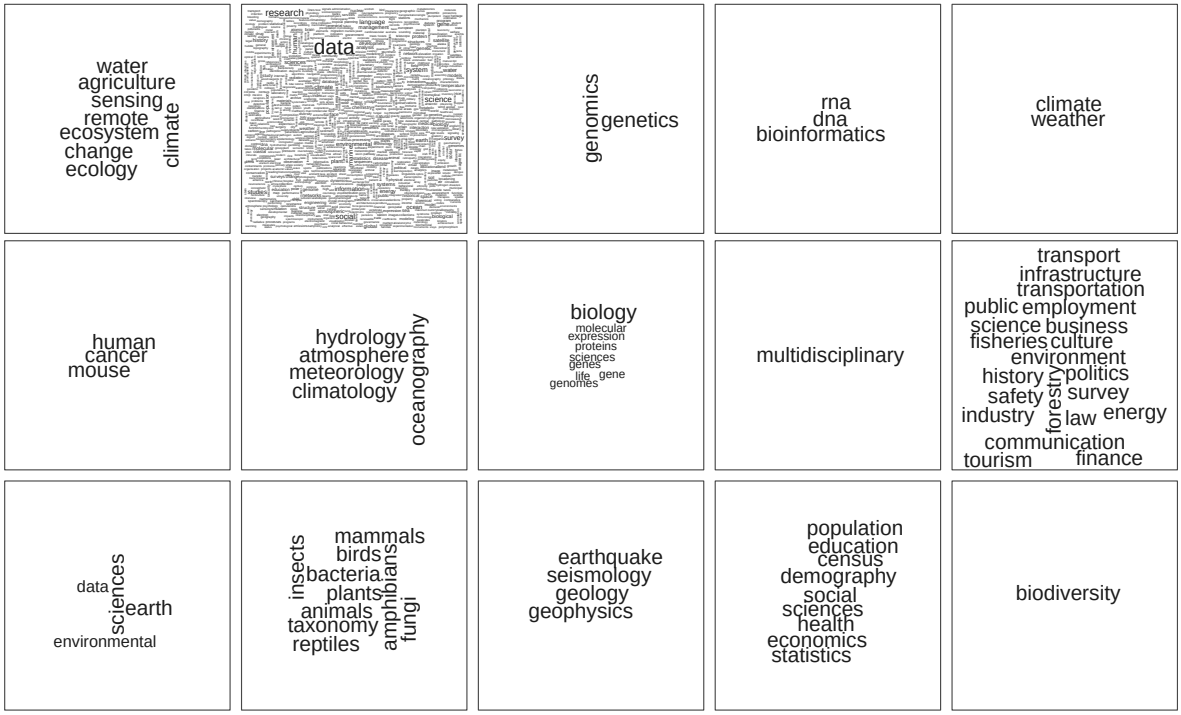
# GRAPH ALGORITHMS

- Recommendation systems
- Pattern detection in coding
- Support for API development and data error detection

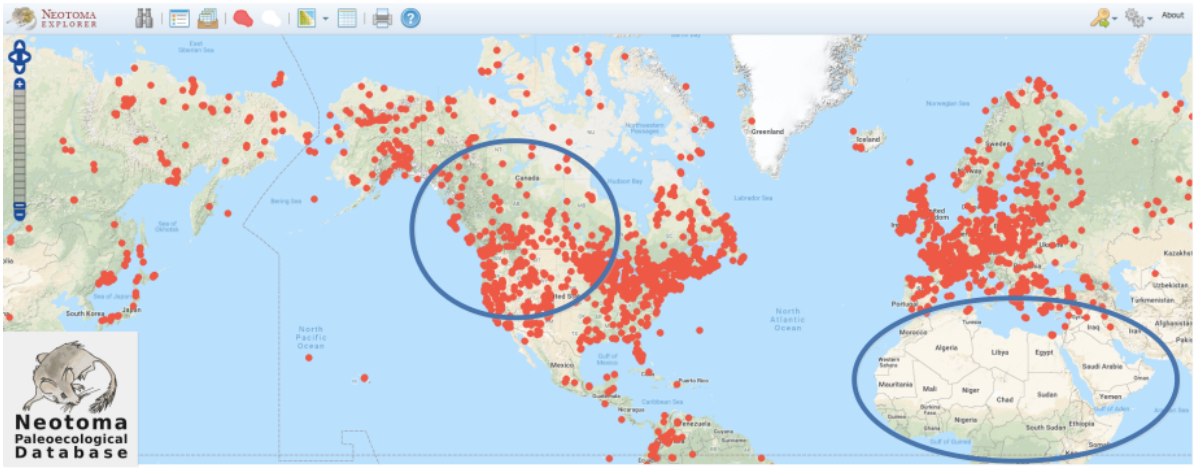
```
results <- neotoma::get_datasets(
  bbox = c(-120, 30, -113, 60))

# data errors
drop_records = c(10, 12, 13, 14)
```

# DEEPER INSIGHTS INTO PATTERNS



# SUPPORTING DATA ACQUISITION




Grimm (1998): "We have inventoried more than 2000 sites that have been studied and are therefore potentially available for inclusion within [Neotoma]"

# SUPPORTING DATA ACQUISITION

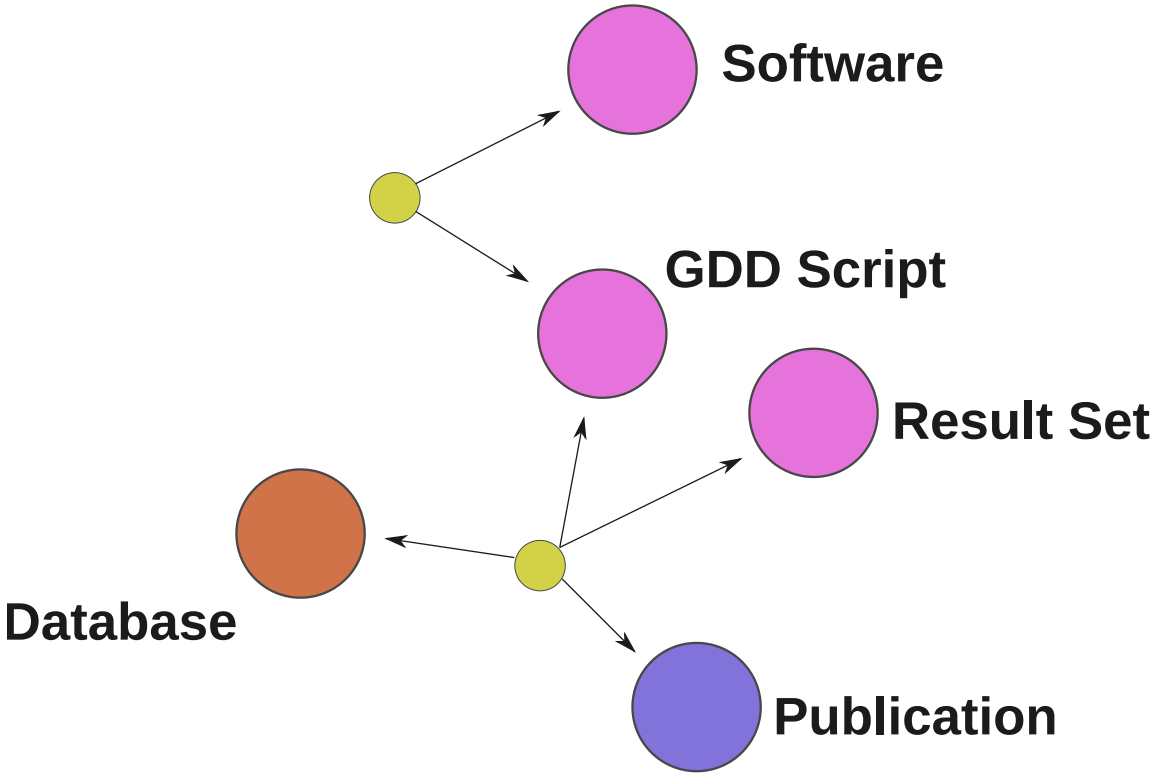
{Quaternary Research 58 , 130 -- 138 ( 2002 ) doi :10.1006 / qres .2002.2353 Paleoenvironmental Changes in the Semiarid Coast of Chile ( ∼ 32 ◦ S ) during the Last 6200 cal Years Inferred from a Swamp -- Forest Pollen Record Antonio Maldonado1 and Carolina Villagra ´ n Laboratorio de Palinolog ´ ıa , Departamento de Biolog ´ ıa , Facultad de Ciencias , Universidad de Chile , Casilla 653 , Santiago , Chile E-mail : amaldona@icaro.dic.uchile.cl Received October 3 , 2001 ; published online August 22 , 2002 Pollen analysis of two sediment records from a coastal swamp forest site in the Chilean semiarid region ( 31 ◦ 50 S ; 71 ◦ 28 W ) shows an alternation of dry and wet phases during the past ∼ 6100 cal yr B.P. .}

# SUPPORTING DATA ACQUISITION

{Quaternary Research 58 , 130 -- 138 ( 2002 ) doi :10.1006 / qres .2002.2353 Paleoenvironmental Changes in the Semiarid Coast of Chile ( ∼ 32 ◦ S ) during the Last 6200 cal Years Inferred from a Swamp -- Forest Pollen Record Antonio Maldonado1 and Carolina Villagra ´ n Laboratorio de Palinolog ´ ıa , Departamento de Biolog ´ ıa , Facultad de Ciencias , Universidad de Chile , Casilla 653 , Santiago , Chile E-mail : amaldona@icaro.dic.uchile.cl Received October 3 , 2001 ; published online August 22 , 2002

**Pollen analysis** of two sediment records from a coastal swamp forest site in the Chilean semiarid region **( 31 ◦ 50 S ; 71 ◦ 28 W )** shows an alternation of dry and wet phases during the past ∼ 6100 cal yr B.P. .}

# SUPPORTING DATA ACQUISITION

# DEVELOPMENT

- Entering Year 1 of a 3 year funding cycle
- **http://github.com/throughput-ec**
    - node.js/Express API
    - database ingest scripts
    - metrics and presentations
    - MIT License, Codes of Contributor Conduct
    - DB Snapshot: **http://bit.ly/throughput-shot**

# QUESTIONS


Farley S, Dawson A, Goring SJ, Williams JW. *2018*. Situating ecology as a big data science: Current advances, challenges, and solutions. *BioScience*. **68**:563–576. DOI: **10.1093/biosci/biy068**

Williams JW, Goring SJ, Emile-Geay J, Fils D, Grimm EC, Lehnert K, McKay N, Myrbo A, Noren A, Park-Boush L, Peters S, Singer B, Uhen M. 2018. Cyberinfrastructure in the Paleosciences: Mobilizing Long-Tail Data, Building Distributed Community Infrastructure, Empowering Individual Geoscientists. **Open Access**

Goring SJ, Graham R, Loeffler S, Myrbo A, Oliver JS, Ormond C, & Williams JW. 2018. *The Neotoma Paleoecology Database: A Research Outreach Nexus*. Elements of Paleontology. Cambridge: Cambridge University Press. [DOI: **10.1017/9781108681582**]

Goring S, Weathers K, Dodds W, Cheruvelil K, Kominoski J, Rüegg J, Sweet L, Utz R. (2014) The collaborative culture of Macrosystems Ecology: Optimizing participant benefits. *Frontiers in Ecology and the Environment*. **12**:39-47. [**Open Access**]



@sjgoring **http://goring.org**