Implications of the Data-Centric Nature of Climate Science for AI & ML

Seth McGinnis, NCAR mcginnis@ucar.edu
Lisa Lloyd, Indiana University; Greg Lusk, Michigan State; Stu Gluck, Johns Hopkins

Abstract

Climate science prioritizes the production and dissemination of data to enhance its value as evidence. The re-use of data in this way depends on how it is packaged. A comparison of the influence of Big Data in biology versus climate science reveals potential hazards associated with the categorization of phenomena. To avoid undesirably constraining downstream research, the development of ontologies and training datasets for machine learning needs to be an open community effort.

Climate Science is Data-Centric

Data-Centric Science prioritizes production and dissemination of data to enhance its value as evidence.

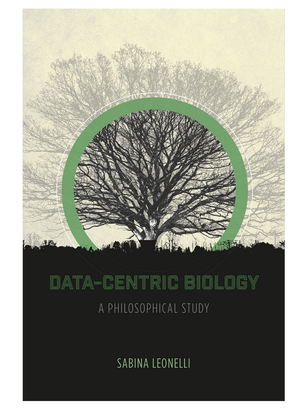
I.e., creation of datasets for use beyond a single experiment – re-using data in new contexts

Climate data is normally shared. Possible reasons why:

- Observations are unique
- Simulations are expensive
- Earth is common to everyone

Above: logos of programs and organizations dedicated to sharing climate data for re-use

We compared Big Data in biology vs climate science



"Data-Centric Biology" by Sabina Leonelli:

an analysis of how Big Data has affected the field of biology.

Climate science also has Big Data, so we wrote a paper.

Big Data

- Volume, Variety, Velocity
- Can't use traditional tools and methods

Biology: Big Variety

- Gene sequences
- Biochemical assays
- Experimental measurements
- Field observations
- Etc.

Focus: Model organisms

Climate: Big Volume

- GCM / RCM outputs
- 3-D, high-freq, ensembles

Focus: Earth system (mostly atmosphere)

Data re-use depends on data packaging

Biology is integrative

Data stored in relational databases

Built on ontologies linking categories:

species, gene, organ, disease, pathogen, etc.…

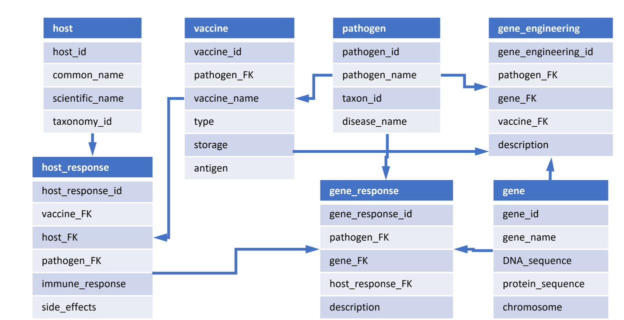
Climate science is distributive

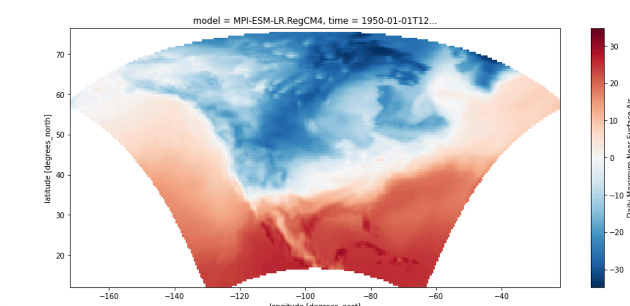
Data stored in flat files

Arrays of values located in space and time

NOT classified into categories

- No catalogs of fronts, storms, etc.
- Just time-varying spatial fields: temperature, velocity, humidity…





Ontology: a description of domain-specific categories and concepts and the relationships between them

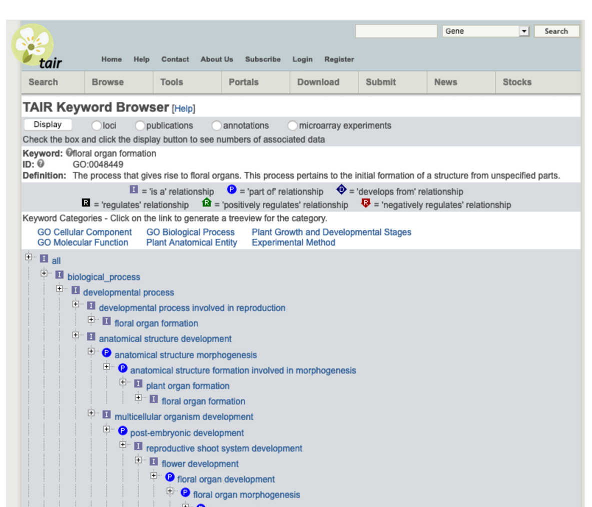
Organization of Big Data using ontologies can constrain downstream research

Example: silvergrass biofuels research – how does oil-boosting gene modification affect flower formation?

miscanthus giganteus is too big for labwork;
use well-studied *miscanthus arabidopsis* instead

DB schema integrates research across topics, but also affects askable questions

- What's represented by tables?
- What's missing from the database?
- What's allowed as a valid entry?



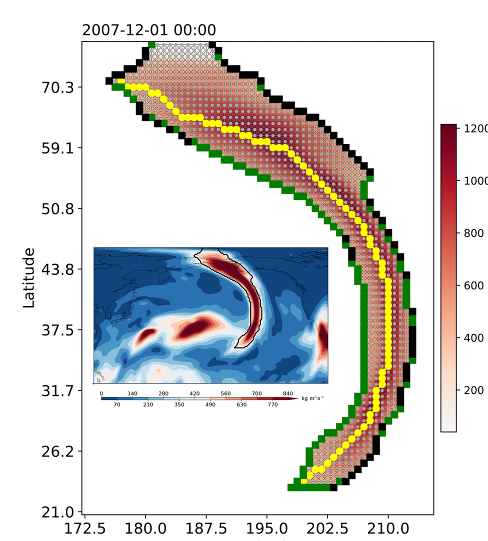
TAIR website for *m. arabidopsis* data

Influence of ontology:

- Data curation matters A LOT
- Privileges easily-digitized data
- Privileges data from well-funded labs with good IT departments

Climate science has avoided these problems so far, but they may become an issue with Machine Learning & AI

Machine Learning is good at problems you solve by looking at pictures, e.g., where's the atmospheric river?



Allows automated cataloging of phenomena in large datasets

Requires large training datasets, which are databases of classified and categorized phenomena like those used in biology

We need to learn lessons from biology about problems caused by ontologization.

In a database:

- Who defines phenomena?
- What counts as a drought / heatwave / atmospheric river / etc?
- Do we have good representation from different geographic regions?
- Do we have input from different communities, esp. downstream users?
- What datasets are used to train algorithms?
- Can the ontology be updated?

Getting this wrong can significantly hinder future research

CONCLUSIONS

Ontology development needs to be an open community effort

We need conferences and workshops to develop community consensus about concepts used in data ontologies.

Ontologies need to be reasonable and responsive.

Definitions need to be regularly reviewed and updated.

Use ARTMIP as an example of how to do things right

- Test a variety of different definitions
- Use a variety of different algorithms and techniques
- Make training data available
- Avoid naïve "gee-whiz" approach to ML (no cowboy coding)

ACKNOWLEDGMENTS

Logos © respective organizations; Book cover © University of Chicago Press; Figure: CC-BY-4.0 Xu et al, https://doi.org/10.5194/gmd-13-4639-2020
This presentation is based on the journal article: Lloyd, Elisabeth A., Greg Lusk, Stuart M. Gluck, and Seth McGinnis. "Varieties of Data-Centric Science: Regional Climate Modeling and Model Organism Research." Forthcoming in *Philosophy of Science.*

WATCH THE VIDEO Portions of this poster were presented in a talk at NCAR's Improving Scientific Software conference in March 2021, which can be seen here: https://youtu.be/IKmYXbRX0eM?t=652