Computing, Data, and Cyberinfrastructure for a Systems Approach to Studying the Earth

Perspectives from the Hydrosphere

Jeffery S. Horsburgh

Associate Professor

Utah State University®

CIVIL AND ENVIRONMENTAL ENGINEERING

The National Academies of Sciences Engineering and Medicine

Advancing a Systems Approach to Studying the Earth: A Strategy for the National Science Foundation

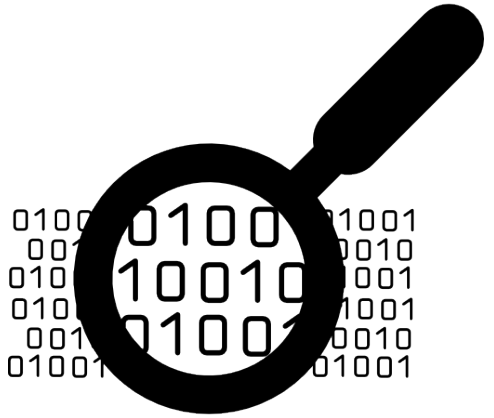
February 4, 2021

Overarching principles

1. Data will permeate through all parts of a systems approach to studying the Earth
   - Data cyberinfrastructure is essential

2. Integrated research across disciplines requires collaboration
   - Data and model interoperability

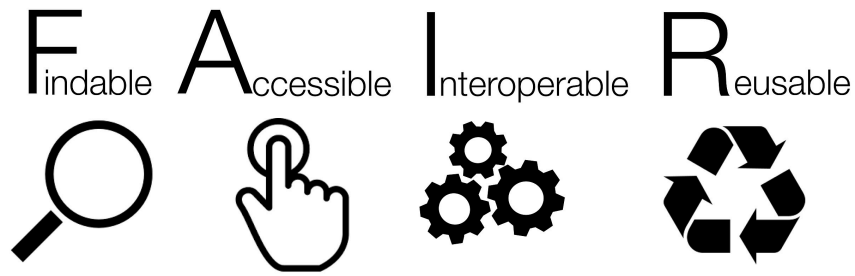
3. We can do more if we can build on existing foundations
   - Reproducibility and composability



A systems approach to studying the Earth requires data

Inputs, outputs, products

Guiding Principles for Scientific Data Management



- Data are first class research products!
- Funders and publishers are now requiring open data sharing

- **Findable**: Data have sufficient metadata and a unique, persistent identifier making data discoverable on the Web
- **Accessible**: Metadata and data are understandable to humans and machines and are available via a trusted repository
- **Interoperable**: Metadata use formal community standards
- **Reusable**: Data have clear metadata, usage license, and information about provenance

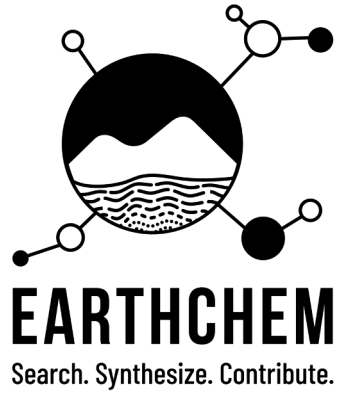
Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, https://doi.org/10.1038/sdata.2016.18.

New Opportunities for Data Sharing and Preservation

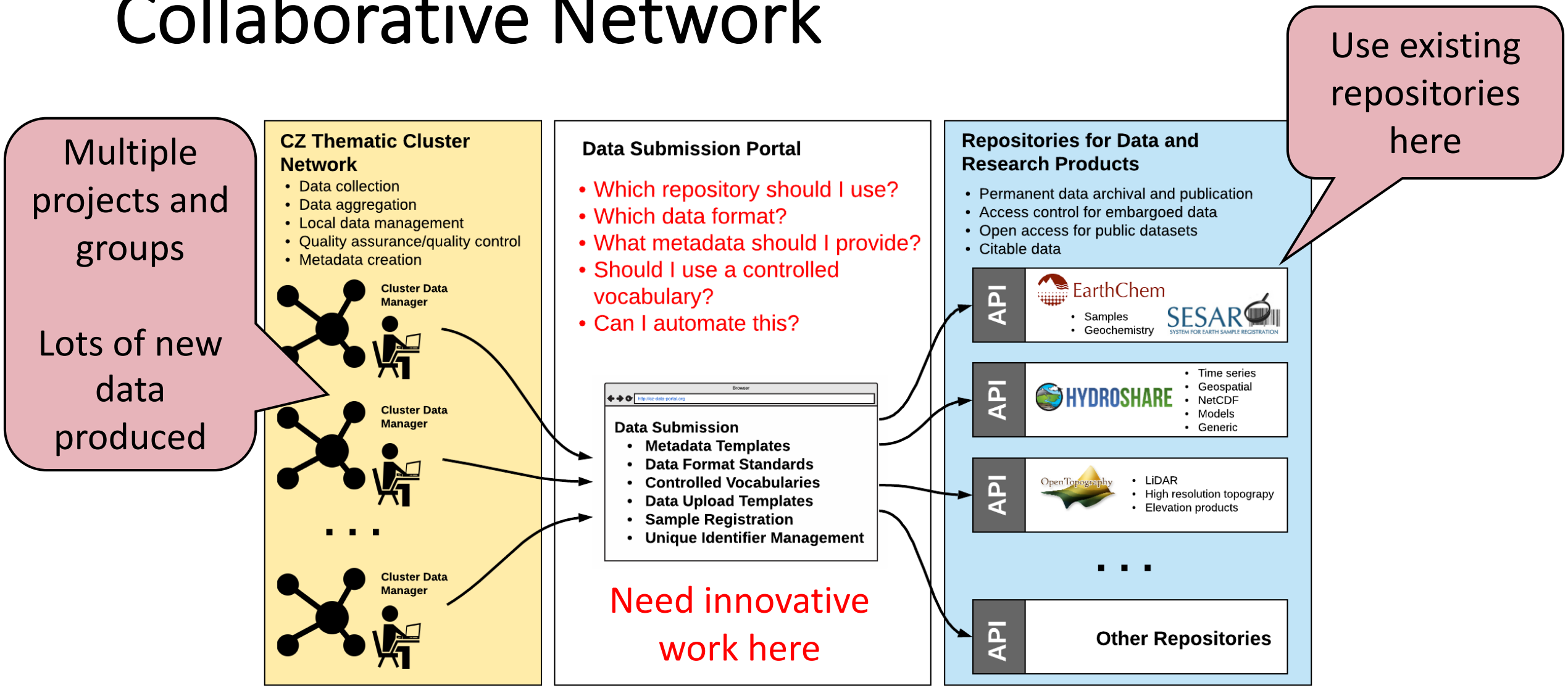
- Growing landscape of data repositories
- Functionality for archival/preservation
- Many are still very much discipline specific
- Registry of Research Data Repositories now lists over 2000 repositories (re3data.org)
- Contain the corpus of Earth science research data

PANGAEA.
Data Publisher for Earth & Environmental Science

CUAHSI
HIS
Sharing hydrologic data



Building a CI for NSF's Critical Zone Collaborative Network



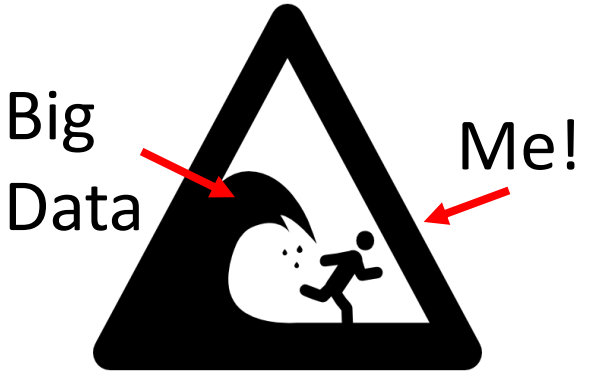
Collaborative Investigator Data Workflow

- Easily create a digital instance of a dataset (or model)
- Quickly share it with colleagues (perhaps privately at first)
- Add value through study, collaboration, annotation, and iteration
- Describe with metadata
- Eventually…share publicly or formally Publish (FAIR!)

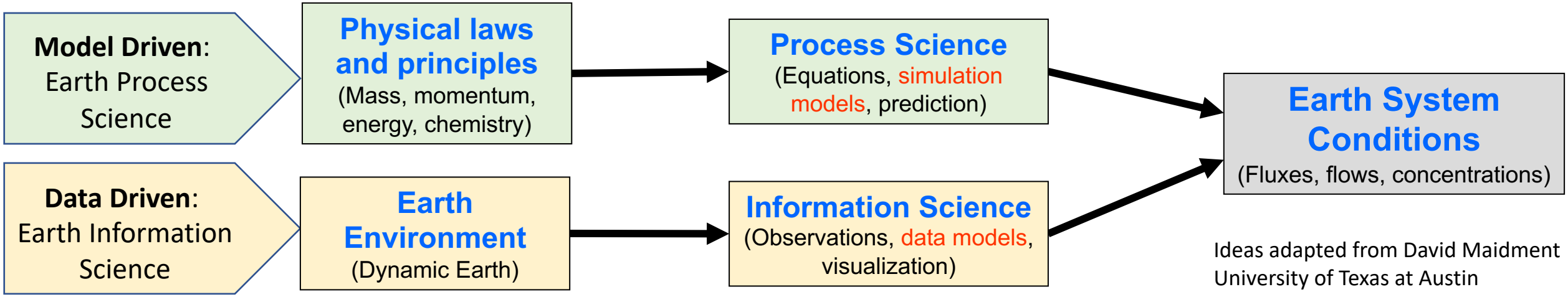
What is the role of data repositories in this scientific workflow?

Big Data

- Earth science big data represent challenges and opportunities
  - Extensively used
  - Too big to easily move around – collaboration looks different
  - How to mine for information?
  - How to simplify size and dimension of data?



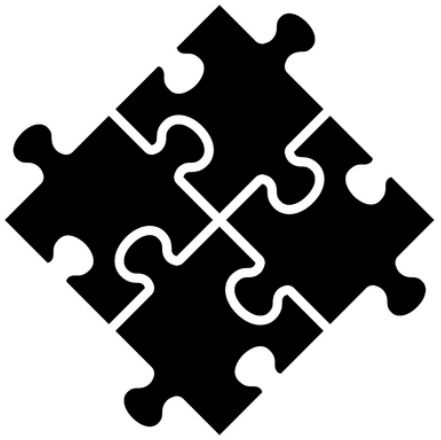
How and when to move from “model driven” research to “data driven” research?



Ideas adapted from David Maidment
University of Texas at Austin

Summary of Data Challenges

- Existing repositories getting better, but how to choose the right repository for data?
- How to share data in a way that it can be understood by users?
- FAIR is not easy - the extent to which data are FAIR affects their value and extent of reuse, FAIR requires committment
- What are the roles that repositories should play in the research process?
- Big Data – what to keep, where to keep it, (how and when) can we shrink it?
- Long term sustainability of research data repositories



A systems approach demands interoperability

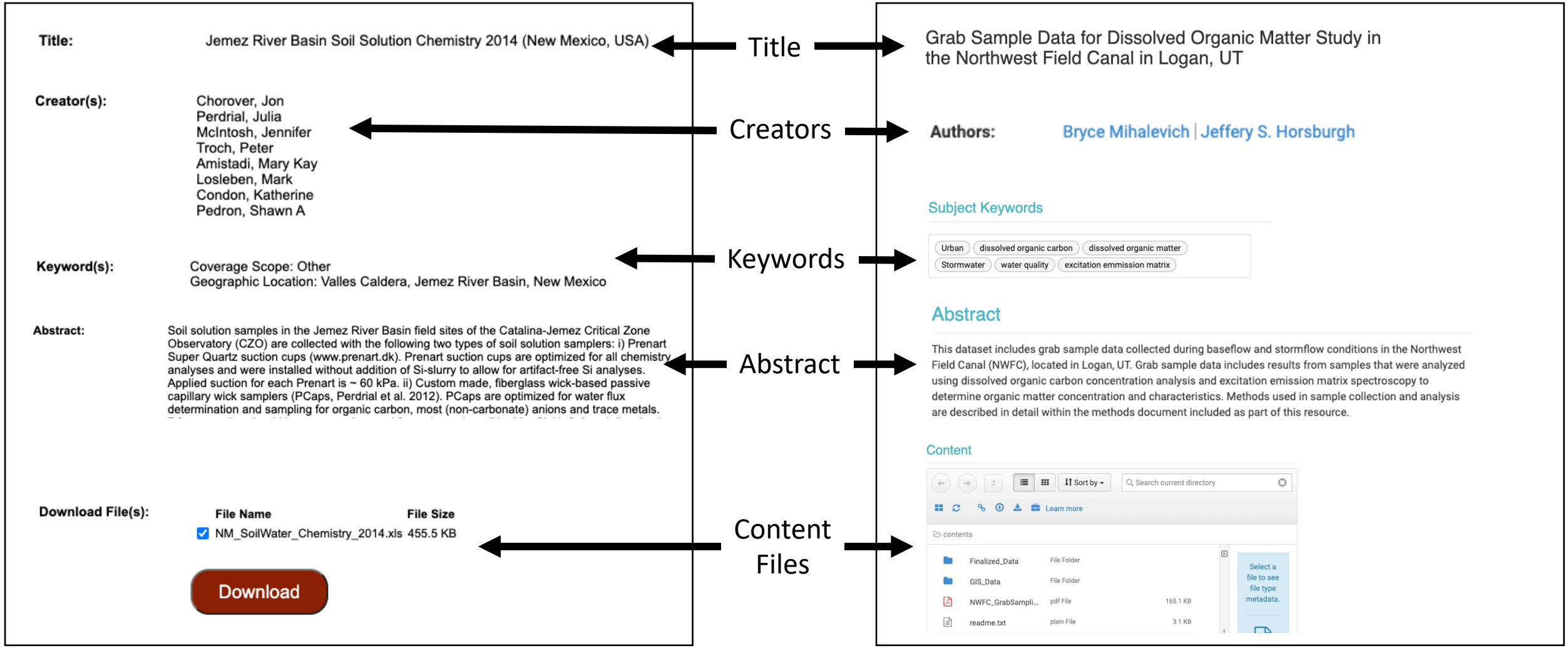
Knowledge is generated through the integration of information from multiple sources

Earth System CI must address this

Repository Interoperability

Chorover, J., Perdrial, J., McIntosh, J., Troch, P., Amistadi, M., Losleben, M., Condon, K., Pedron, S. 2018. Jemez River Basin Soil Solution Chemistry 2014 (New Mexico, USA), Version 1.0. Interdisciplinary Earth Data Alliance (IEDA). https://doi.org/10.1594/IEDA/111144. Accessed 2021-02-02.

Mihalevich, B., J. S. Horsburgh (2017). Grab Sample Data for Dissolved Organic Matter Study in the Northwest Field Canal in Logan, UT, HydroShare, https://doi.org/10.4211/hs.a3a9ba772aac4cba9533b35bb6b5fe42

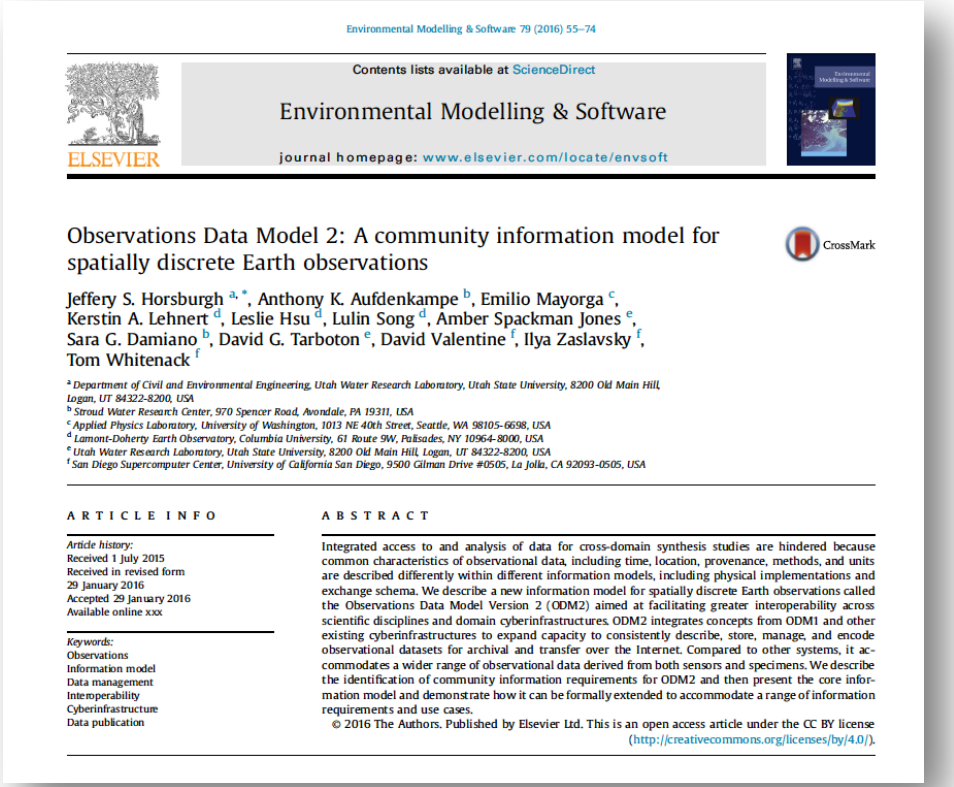


Data Integration is Difficult

- Many research scenarios require integration of multiple data types across different Earth science domains
- Data from multiple repositories use different format, syntax, and semantics
- Common characteristics of observational data (time, location, provenance, methods, units) are described using different constructs within different systems

***Example**: Understanding a soil profile's geochemical response to extreme weather events requires integration of hydrologic and atmospheric time series with geochemical data from soil sample fractions collected over various depth intervals from soil cores or pits at different positions on a landscape*

Standard information models can help



Environmental Modelling & Software 79 (2016) 55–74

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Observations Data Model 2: A community information model for spatially discrete Earth observations

Jeffery S. Horsburgh [a, *], Anthony K. Aufdenkampe [b], Emilio Mayorga [c], Kerstin A. Lehnert [d], Leslie Hsu [d], Lulin Song [d], Amber Spackman Jones [e], Sara G. Damiano [b], David G. Tarboton [e], David Valentine [f], Ilya Zaslavsky [f], Tom Whitenack [f]

[a] Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
[b] Stroud Water Research Center, 970 Spencer Road, Avondale, PA 19311, USA
[c] Applied Physics Laboratory, University of Washington, 1013 NE 40th Street, Seattle, WA 98105-6698, USA
[d] Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, NY 10964-8000, USA
[e] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
[f] San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive #0505, La Jolla, CA 92093-0505, USA

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
29 January 2016
Accepted 29 January 2016
Available online xxx

Keywords:
Observations
Information model
Data management
Interoperability
Cyberinfrastructure
Data publication

ABSTRACT

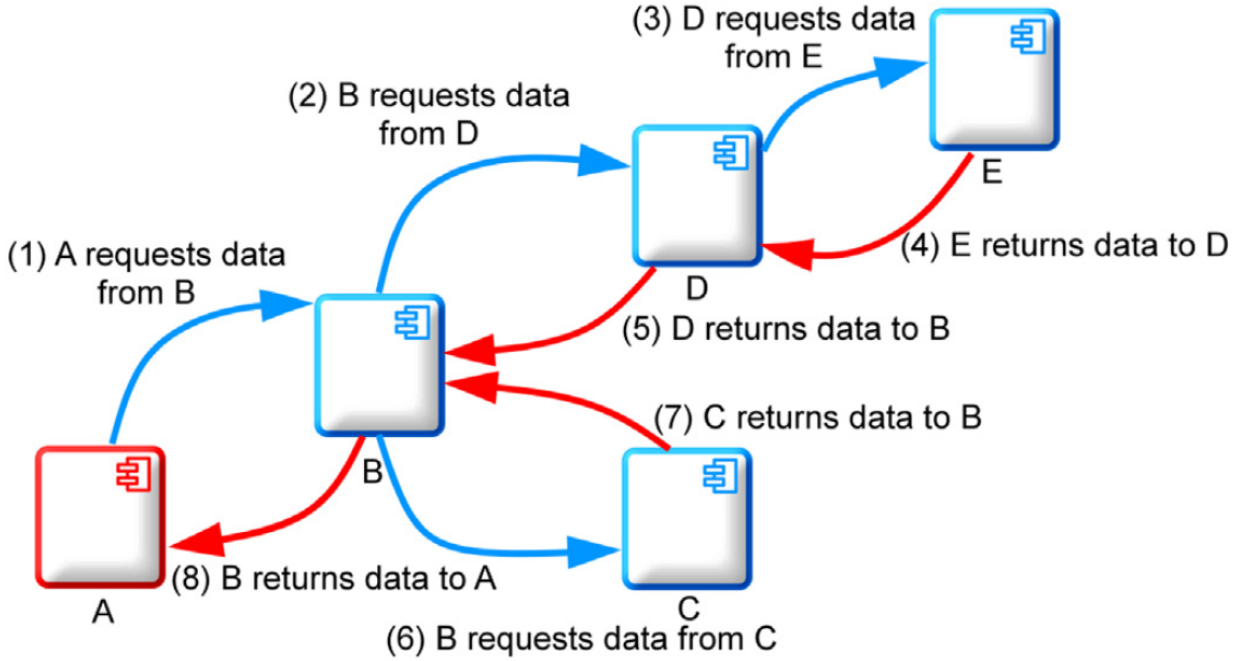
Integrated access to and analysis of data for cross-domain synthesis studies are hindered because common characteristics of observational data, including time, location, provenance, methods, and units are described differently within different information models, including physical implementations and exchange schema. We describe a new information model for spatially discrete Earth observations called the Observations Data Model Version 2 (ODM2) aimed at facilitating greater interoperability across scientific disciplines and domain cyberinfrastructures. ODM2 integrates concepts from ODM1 and other existing cyberinfrastructures to expand capacity to consistently describe, store, manage, and encode observational datasets for archival and transfer over the Internet. Compared to other systems, it accommodates a wider range of observational data derived from both sensors and specimens. We describe the identification of community information requirements for ODM2 and then present the core information model and demonstrate how it can be formally extended to accommodate a range of information requirements and use cases.

Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., Spackman Jones, A., Damiano, S. G., Tarboton, D. G., Valentine, D., Zaslavsky, I., Whitenack, T. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations, *Environmental Modelling & Software*, 79, 55-74, http://dx.doi.org/10.1016/j.envsoft.2016.01.010

Model Compatibility and Integration

Transdisciplinary Earth science often requires integration of models from multiple disciplines

- Model coupling and model integration tools have advanced (OpenMI, ESMF, CSDMS, BMI, etc.)
- Multiple model coupling approaches (tight, loose, data centric)
- Space and time scales are not always well matched
- Semantics at the interfaces between components



OpenMI "request and reply" data exchange mechanism. Component A is the controller/trigger for the simulation.

Buahin, C. A., Horsburgh, J. S. (2018). Advancing the Open Modeling Interface (OpenMI) for integrated water resources modeling, Environmental Modelling & Software, 108, 133-153, https://doi.org/10.1016/j.envsoft.2018.07.015.

Summary of Interoperability Challenges

- Opportunities for interoperability across repositories
  - Metadata
  - Formats/encodings
  - Vocabularies/semantics
  - Packaging and delivery
  - Designing for diversity in data
- Facilitating data integration
  - Standard information models could be a preventative
  - “Data munging” treats the symptoms of the disease
- Modeling
  - How to resolve mismatches in space, time, and semantics (scale and interpretation)
  - Which model coupling approach is effective?

Reproducibility is key

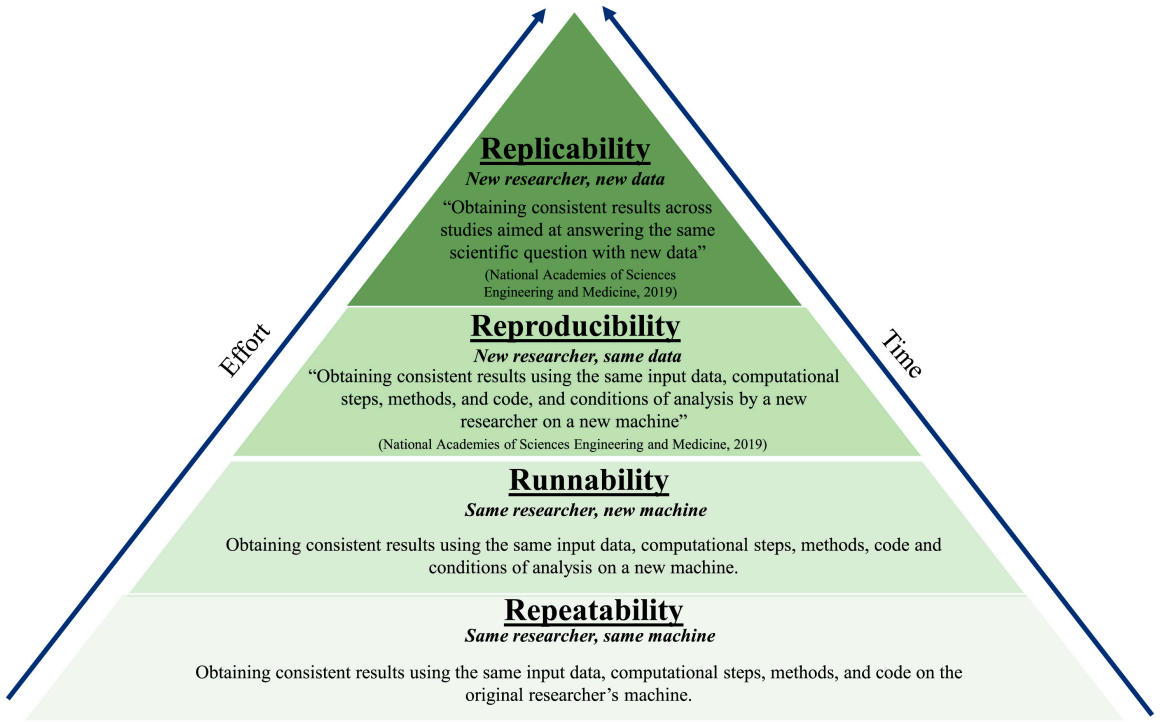
“If I have seen further it is by standing on the shoulders of Giants.”

Isaac Newton, 1625

Building trust in research requires transparency and reproducibility

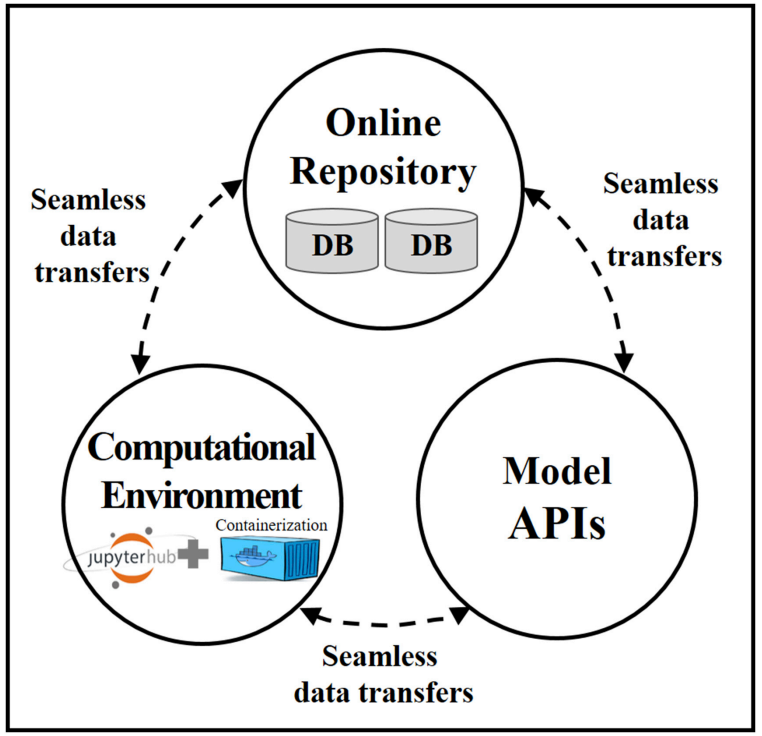
How can CI support reproducibility?

What it means to reproducible



Essawy, B., J. Goodall, D. Voce, M. Morsy, J. Sadler, Y. D. Choi, D. Tarboton and T. Malik (2020), A taxonomy for reproducible and replicable research in environmental modeling, *Environmental Modelling & Software*, 134:104753, https://doi.org/10.1016/j.envsoft.2020.104753

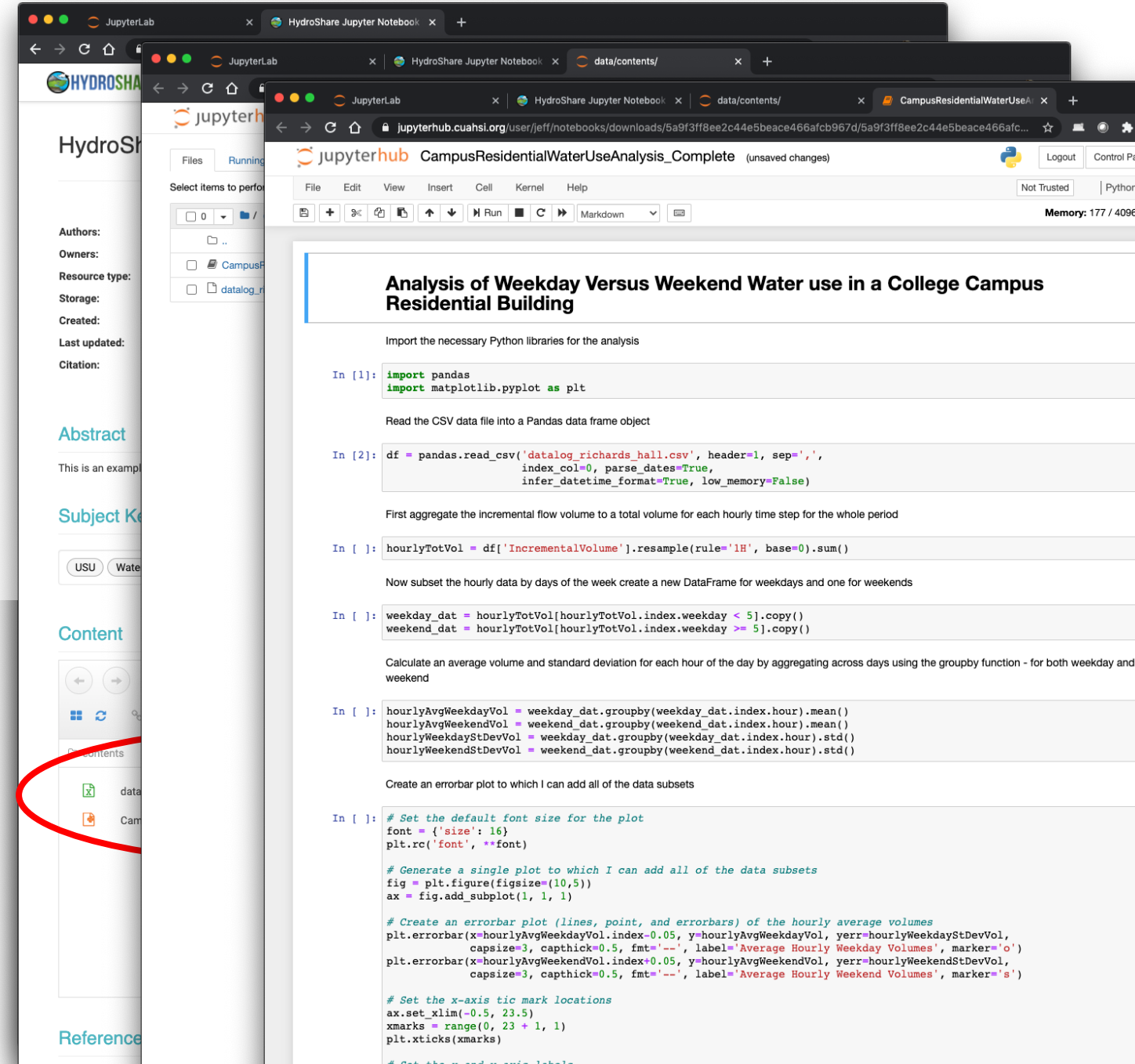
How CI can help (modeling context)



Choi, Y. D., J. L. Goodall, J. M. Sadler, A. M. Castronova, A. Bennett, Z. Li, B. Nijssen, S. Wang, M. P. Clark, D. P. Ames, J. S. Horsburgh, H. Yi, C. Bandaragoda, M. Seul, R. Hooper and D. G. Tarboton (2021), Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model application programming interfaces, Environmental Modelling & Software, 135:104888, https://doi.org/10.1016/j.envsoft.2020.104888

Creating and Sharing Reproducible Analyses

- Reproducible analyses: Sharing data and code together in a repository
- Linking repositories with computational environments
- Repositories as a gateway to high performance computing and cloud services



Summary of Reproducibility Challenges

- What does it mean for computational science to be reproducible?
- How to best link repositories to computation/execution environments?
- How to build shared access to data and computation?
- How to promote more consistent data workflows and data reuse?
- How to provide the “right” computational environment (e.g., in a JupyterHub) and how to maintain it over time?
- How to overcome platform and library dependencies?

A systems approach must bridge “digital divides”

The gaps between those who have programming, data science, and computational skills and those who do not

Challenges and Trends in Environmental Data Science

- Shortage of trained experts
  - Engineering and science programs struggle to fit “data science” into their curriculum
  - Programming and computational skills are often self taught
  - Begs for changing our approach to education and workforce development
  - Collaborative teams can build the needed skillsets
- Methodological gaps for real applications
  - Lack of guidance for mapping methods to applications
  - Methods for choosing the right data
  - Need for development of new methods

Environmental Modelling & Software 106 (2018) 4–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Environmental Data Science

Karina Gibert [a, *], Jeffery S. Horsburgh [b], Ioannis N. Athanasiadis [c], Geoff Holmes [d]

[a] Dep. Statistics and Operations Research, Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center (KEMLG at IDEAI), Research Institute on Science and Technology for Sustainability, Universitat Politècnica de Catalunya-BarcelonaTech, Spain
[b] Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-8200, USA
[c] Information Technology Group, Wageningen University, The Netherlands
[d] Department of Computer Science, University of Waikato, New Zealand

ARTICLE INFO

Article history:
Received 14 February 2018
Received in revised form
11 April 2018
Accepted 24 April 2018

Keywords:
Data Science
Environmental science
Data driven modelling

ABSTRACT

Environmental data are growing in complexity, size, and resolution. Addressing the types of large, multidisciplinary problems faced by today's environmental scientists requires the ability to leverage available data and information to inform decision making. Successfully synthesizing heterogeneous data from multiple sources to support holistic analyses and extraction of new knowledge requires application of Data Science. In this paper, we present the origins and a brief history of Data Science. We revisit prior efforts to define Data Science and provide a more modern, working definition. We describe the new professional profile of a data scientist and new and emerging applications of Data Science within Environmental Sciences. We conclude with a discussion of current challenges for Environmental Data Science and suggest a path forward.

1. Introduction

“Data Science is the science of dealing with data …” (Naur, 1974)

In recent years, we have observed an increasing popularity of Data Science methods that seem to be in the focus of many organizations, including those interested in a better comprehension or management of environmental systems. Data Science is already widely used in business to design successful strategies and policies, and the economic sector is facing a significant transformation as a result of the penetration of data-driven innovation in the business core. We believe that a similar transformation is underway within many scientific disciplines, among them those within the Environmental Sciences, to investigate the benefits that can be realized through use of appropriate Data Science approaches.

In this paper, we analyze the origins of Data Science as a new discipline that is diverse enough to be applied to any domain, including those within the Environmental Sciences. The potential of Data Science to advance our knowledge of the laws governing complex environmental phenomena is enormous. The technological development requisite for collecting the volume and resolution of data required to study these phenomena is mature, but classical data analysis methods are, in many cases, insufficient to cope with the size, speed and diversity of information sources providing evidence under the variety of forms (text, videos, audio recordings, numbers, images) that require global analysis and local tuning to elicit the hidden, relevant knowledge to support higher level decision making. Many investigators are already investigating how Data Science can address this deficiency.

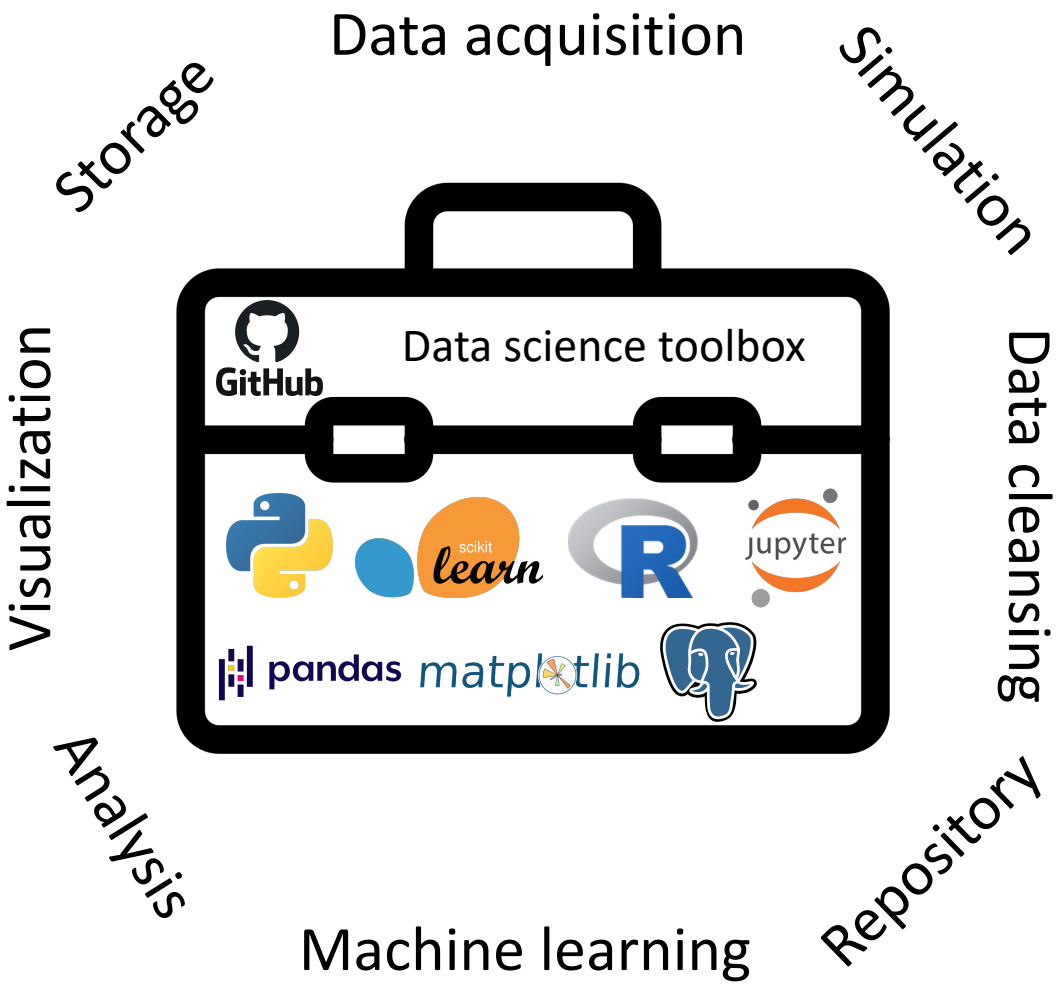
We present the contributions of Data Science, together with an analysis of the new, specific skills associated with its inherent multidisciplinarity. As there is no common definition of Data Science, in the paper we present several definitions that have been used in the past and a propose a new conceptualization of what Data Science means. A discussion is also provided regarding its contact points with other emerging disciplines, such as Big Data Analytics. Emerging opportunities for new applications in Environmental Sciences are described. While not an exhaustive description of the opportunities for Data Science in Environmental Science applications, a wide perspective in the area is provided. Being an emergent field, a number of open issues envisage fertile areas for new research in the near future. The paper also provides some highlights, challenges, and trends with the aim to push the development of the Data Science field in general, and in

* Corresponding author.
E-mail address: karina.gibert@upc.edu (K. Gibert).

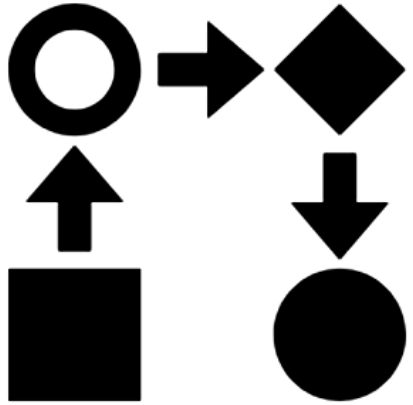
https://doi.org/10.1016/j.envsoft.2018.04.005
1364-8152/

Gibert, K., Horsburgh, J. S., Athanasiadis, I. A., Holmes G. (2018). Environmental Data Science, *Environmental Modelling & Software*, 106, 4-12, https://doi.org/10.1016/j.envsoft.2018.04.005.

Cyberinfrastructure can help bridge



- Provide data management, visualization, and analysis tools that advance scientists' capabilities
- Build them in a way that makes them more accessible



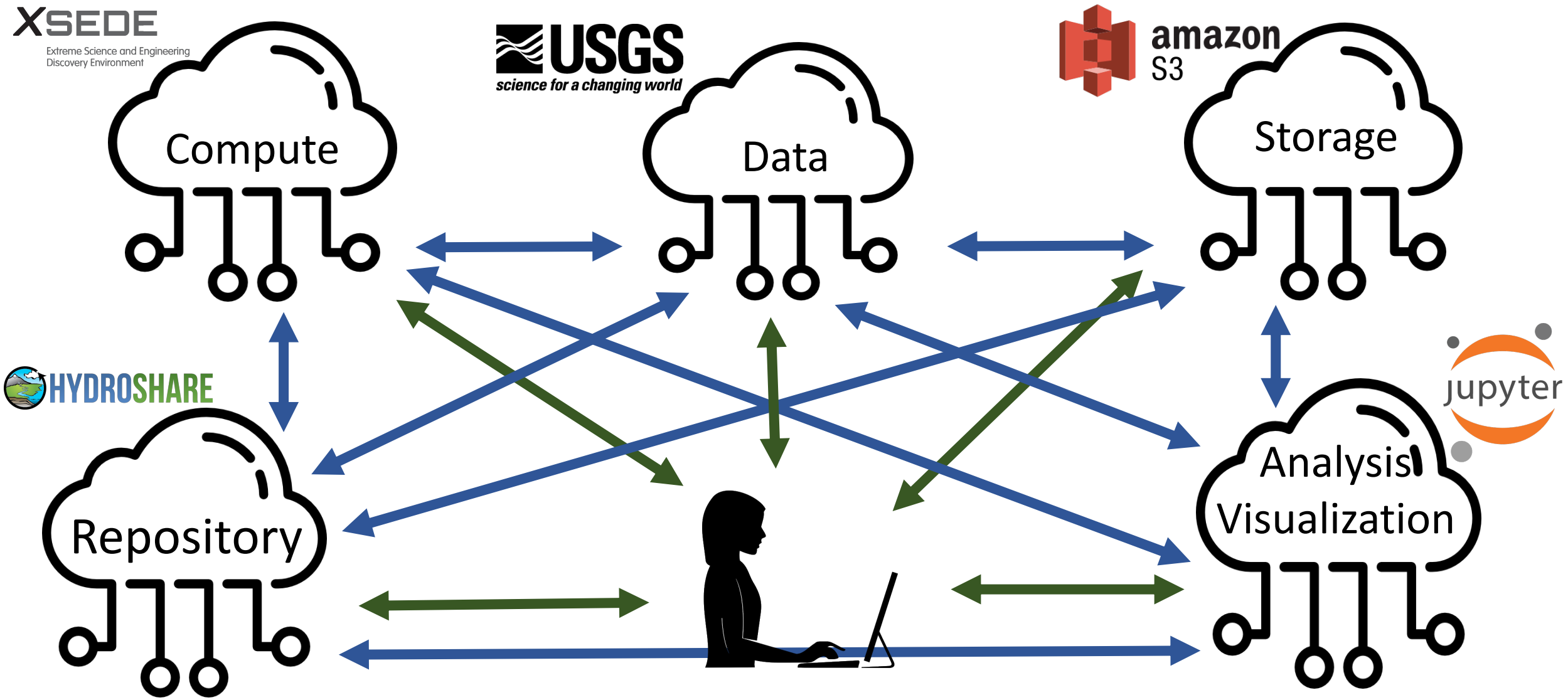
A systems approach could benefit from improved “composability”

How do we connect existing systems?

The old working paradigm

- The “Beefy” Desktop Computer
  - Large hard drive(s)
  - Multiple processors
  - Large amounts of RAM
- Most work done locally
  - Modeling
  - Geospatial analyses
  - Data processing
  - Visualization
- Isolated – can be difficult to share and reproduce

More services are now cloud based
Composing workflows that bridge clouds is hard



Summary of Composability Challenges

- Good cyberinfrastructure components exist
  - We have the parts, but how to connect to better enable Earth science?
- Need for standardized interfaces between systems
  - Standardized web services APIs
  - Client code libraries for popular languages
- Need for standardized packaging and representation of data and computational workflows for transport (data, metadata, provenance)
- Develop architectures that are transparent to users
  - e.g., users collaborate in Google Docs without caring about the how and where. Can Earth systems modeling be this transparent?
- (Continue) supporting changes in the way people work
  - Delivering research and analysis functionality as services over the web
  - We need to be able to string existing pieces together to compose analysis and modeling workflows
  - Enable scientists to more easily share and collaborate around data and analyses

Thank you!

Jeffery S. Horsburgh

jeff.Horsburgh@usu.edu

Utah State University

CIVIL AND ENVIRONMENTAL ENGINEERING

These slides are available in HYDROSHARE

Horsburgh, J. S. (2021). Computing, Data, and Cyberinfrastructure for a Systems Approach to Studying the Earth: Perspectives from the Hydrosphere, HydroShare, http://www.hydroshare.org/resource/331fabb514d641e8b64a832ed464c405