“Preparing for a cloudy future”

Prepared by NOAA’s Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG)

Chair: Chelle Gentemann, Farallon Institute / Earth and Space Research
Molly Jahn, Jahn Research Group
Thomas Huang, NASA/JPL
Shane Glass, Google
Ana Pinheiro Privette, Amazon
Karen Stocks, Ocean Observatories Initiative
Eoin Howlett, Applied Science
Kandis Boyd, NOAA
Lucas Joppa, Microsoft

Outline

This is a brief report from NOAA's Science Advisory Board (SAB) Data Archive and Access Requirements Working Group (DAARWG) on key issues and potential recommendations related to preparing analysis-ready datasets, training researchers to work in the cloud, preparing training data for machine learning, and the process for agile cloud implementation and deployment.

(1) Recommendations for preparing **analysis-ready datasets**.
(2) Recommendation for **training** researchers to work in the cloud.
(3) Recommendations for **preparing training data** for machine learning.
(4) Recommendations regarding process for **agile** cloud implementation and deployment

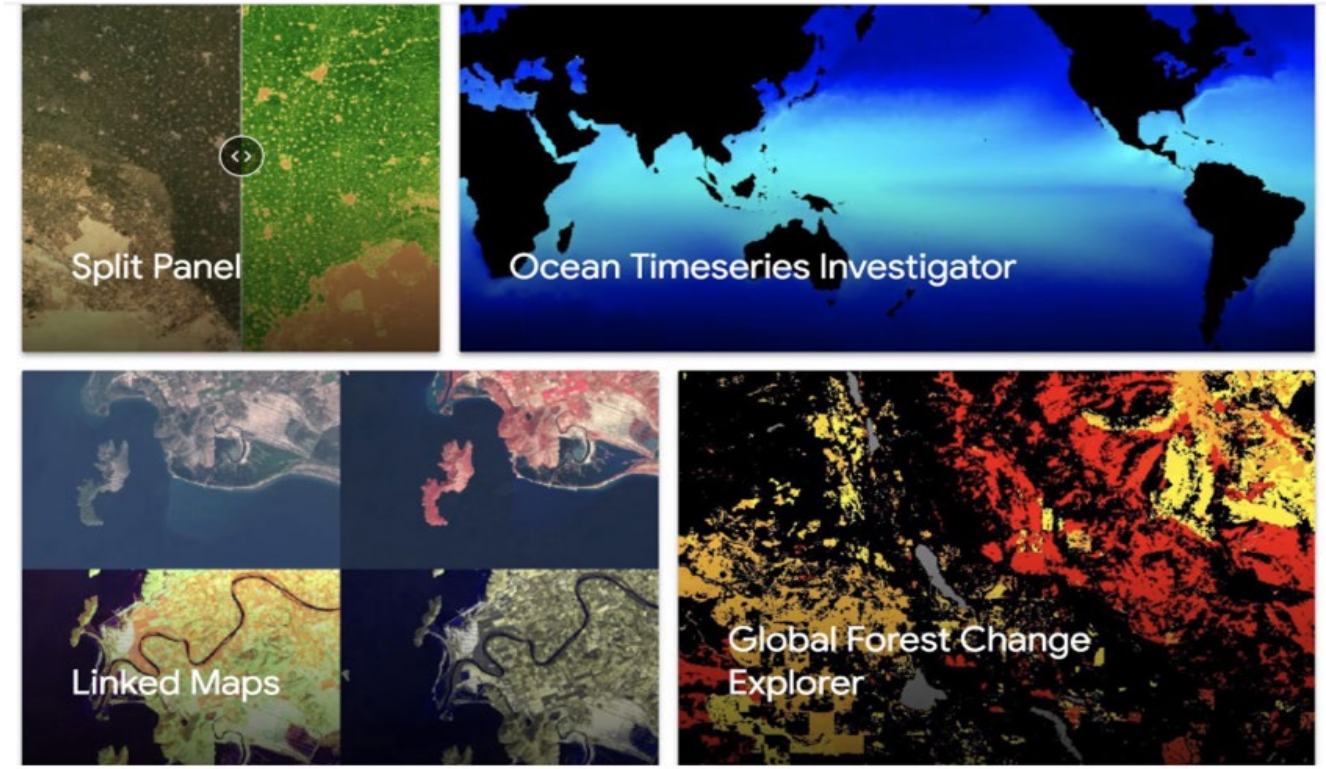
Consortia

Bringing together NASA, NOAA, FEMA data, and that from other Federal agencies, can help us better understand the economic and social impact of disasters, support municipal and state managers, improve the health and safety of American citizens and foster resilient economic development. Moving onto the cloud is inherently **a multi-disciplinary scientific and technological challenge** that will require a coordinating a complex landscape of data, tools, services, science, and personnel.

Recommendation 0.1: Promote the creation of **consortia** focused on specific societal benefits that include private sector, public sector, NGOs and civil society that focus on developing cloud-based solutions using NOAA data.

1. Recommendations for preparing analysis-ready datasets

ARD: Dataset(s) that are processed, organized, described, and accessed to minimize duplicative and/or unnecessary effort from users prior to analysis, enable immediate analysis with minimal user effort, and maximize interoperability with the time-series and with other datasets.



Earth Engine Example Apps from Share your analyses using Earth Engine Apps

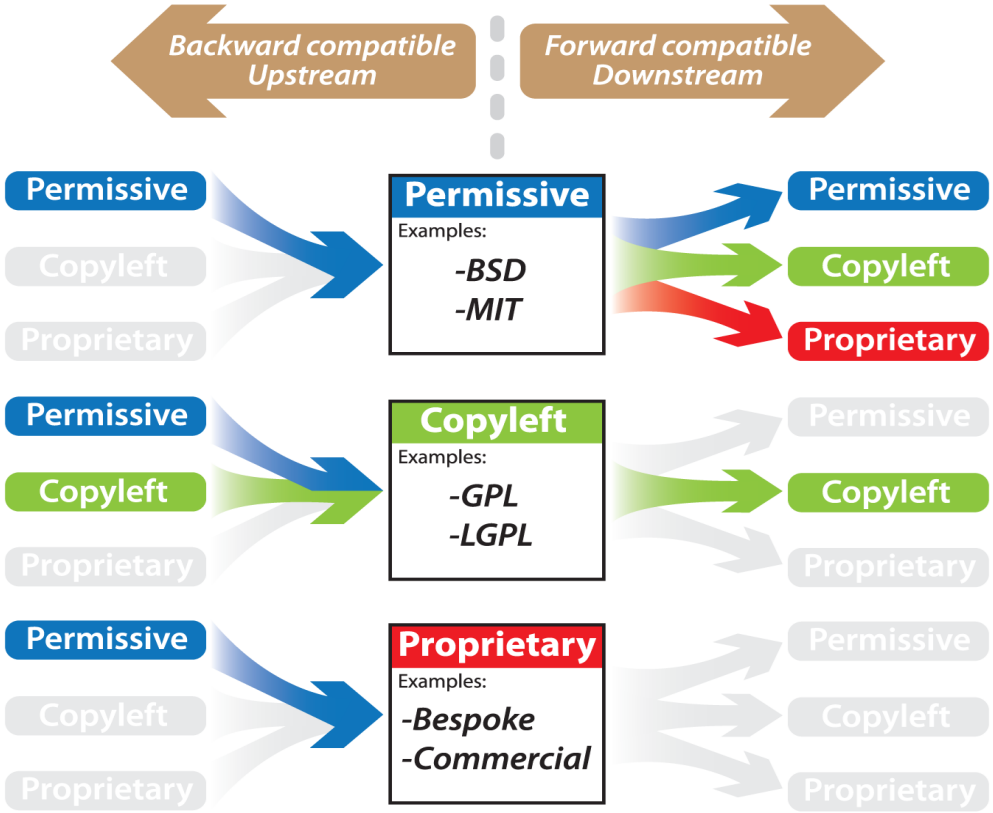
1. Recommendations for preparing analysis-ready datasets

Recommendation 1.1: NOAA's strategy for providing public data **access** should:

a) Account for the needs and **priorities** of all user communities through a systematic and rigorous process, including an analysis of existing data usage patterns;
b) Publicly publish documentation that clearly describes the entire **process** of converting data access to ARDs, including its prioritization criteria and any technical processes utilized;
c) Make available data in its **existing format** in the Cloud as soon as possible to maximize user benefit while their conversion to ARDs is ongoing.

1. Recommendations for preparing analysis-ready datasets

Recommendation 1.2: NOAA should apply a clear, consistent, concise, and **permissive open license** to its data and software. NOAA should present this information along side every dataset at all access points, and include it in the metadata.

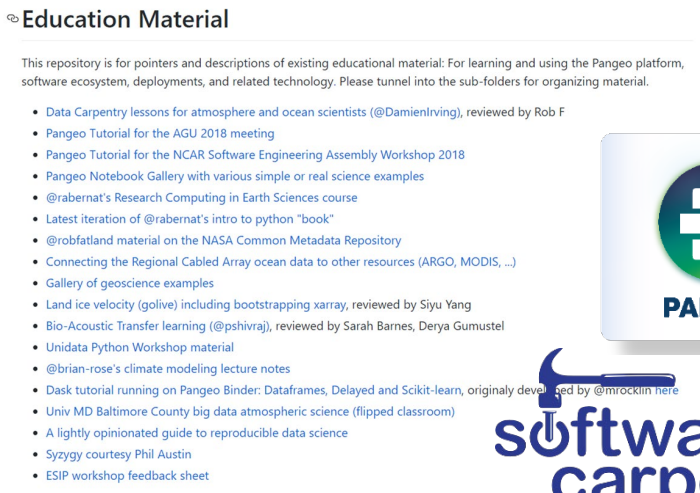


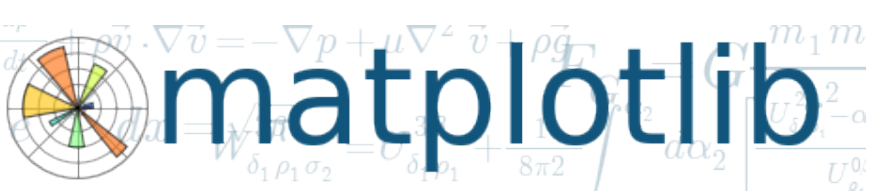
2. Recommendations for training researchers to work in this area

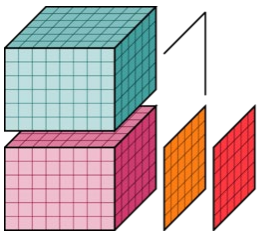
Recommendations 2.1: Offer **trainings** on existing open source technology (open source software libraries, containerization, software stacks, cloud-deployments).

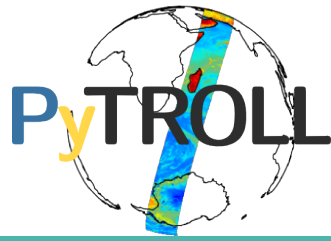
Recommendation 2.2: **Build on existing** open source training resources for on-boarding new collaborators and retraining existing workforce.

Recommendation 2.3: **Support** open source software libraries.







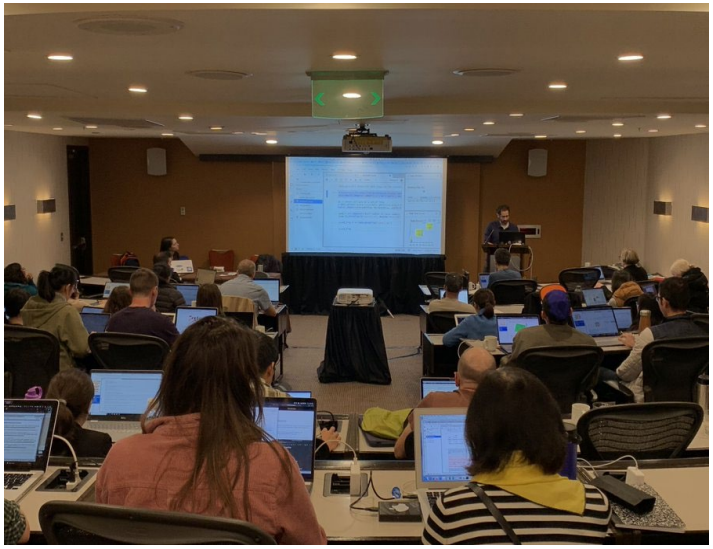


2. Recommendations for training researchers to work in this area

Recommendation 2.4: Develop a mechanism to fund **Hackweek** style events to address specific infrastructure issues or science challenges.

Recommendation 2.5: Ensure inter-agency **communication** on lessons learned to help NOAA's approach to retraining it's workforce.

Recommendation 2.6: Participate in **consortia** with external groups focused on using big data and cloud computing to address societal problems.



Pangeo
tutorial 2019
Fall AGU

3. Recommendations for preparing training data for machine learning

Recommendation 3.1: Develop a NOAA-wide survey to **identify pre-existing** training datasets that may only require minimal reformatting and public distribution.

Recommendation 3.2: Connect external and internal experts to develop and **publish guidance** for data providers to help create clean, quality-control, and labeled ML training data.

Recommendation 3.3: Use the results from R3.1 to identify where there are substantial synergies across training dataset development to develop an **enterprise solution** and minimize redundancies.

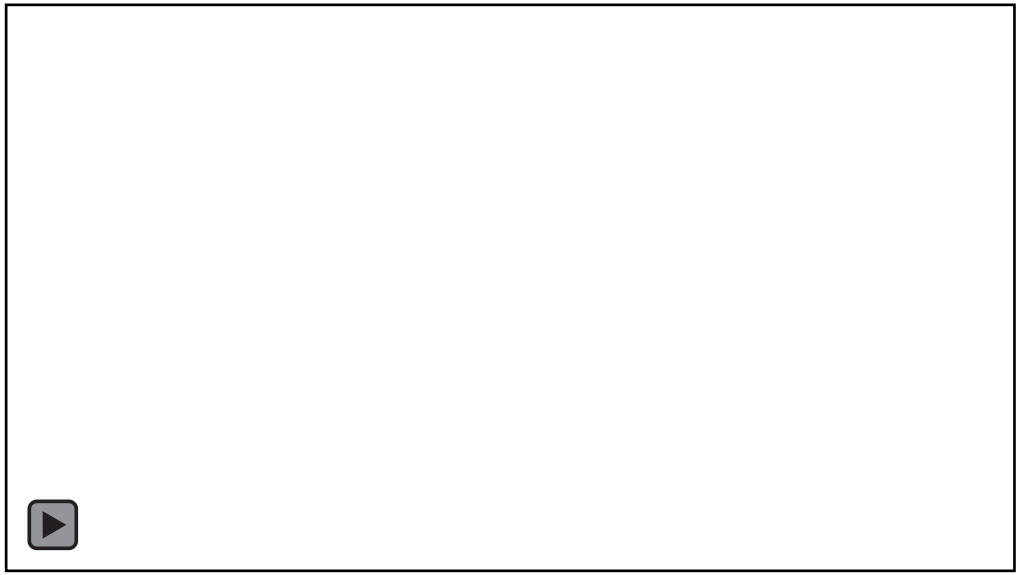
Recommendation 3.4: Publically **share** R3.1 identified datasets in a widely accepted format.

Recommendation 3.5: **Document** the use of these new data.

4. Recommendations regarding process for agile cloud implementation and deployment

Recommendation 4.1: Develop a **categorization** for where, and to what level, more formal software development processes would benefit NOAA.

Recommendation 4.2: When appropriate, NOAA should consider adopting an **agile** software development strategy, including Test-Driven, Continuous Integration (CI), and container-based deployment software.

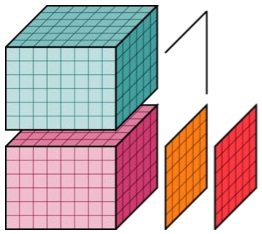


GitHub

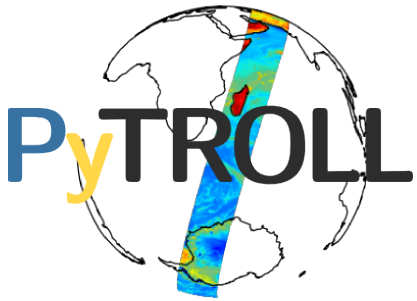
GitLab

DAARWG

Questions?



xarray



GitLab