Data Schools

RDM Lab

Presented by
Steve Diggs (University of California Office of the President)
Shaily Gandhi (PLUS, Salzburg University)

ICTP - Trieste, ITALY: 2024-08-17

Adapted from original material by: S. “Venkat” Venkataraman and NASA TOPS 101

LOGISTICS / TEAMWORK

- All activities will be done in your teams
- Work in teams of 3 or 4
- The slide deck is pre-made and one the blank pages is for your group.
- Every group will have a chance to make a final presentation

Why is Open Science Important?

- Being more open encourages best research practices and makes it easier for you to build on your work.
- Open results have more visibility and impact.
- Open science encourages more collaborative science.

Figure: There are many benefits of open science. CC-BY Danny Kingsley & Sarah Brown

.

Open Science Can Accelerate the Pace of Scientific Progress

Open science practices accelerate the pace of scientific discovery by involving ideas and labor from the broader community.

- **The RAPID RESPONSE to the Covid-19 Pandemic showed Open Science in action to accelerate discovery**
- **CITIZEN SCIENTISTS DRIVE NEW RESEARCH METHODS**
  - https://www.nasa.gov/missions/europa-clipper/citizen-scientists-enhance-new-europa-images-from-nasas-juno/
- **RADAR DATA TO KEEP AN EYE ON CLIMATE CHANGE**
  - https://aws.amazon.com/blogs/publicsector/the-birds-in-the-cloud-how-the-university-of-oklahoma-uses-nexrad-data-to-study-birds/

Group Activity

Think about Open Science (10 min)

In this activity reflect on your answers to the questions and then compare your thoughts with others in the class

- In your field, what steps are being taken to increase openness, and what stands in the way?
- What could help to increase openness?
- What stands in the way?
- What is the Benefits to You ?
- Can you find your own previous work, post-publication and/or pre-publication?
- Can you bring your research materials (data, code, results) with you if you change institutions?
- Can you find the work of your collaborators? Of scientists in other fields that you find interesting?
- Have you reached out to others to collaborate with them after finding interesting results?
- Are people in your field giving and getting credit for work done?

Activity: Think about Open Science (10 min)

(Groups of 3 or 4)

Data Schools

In this activity reflect on your answers to the questions and then compare your thoughts with others in the class

- In your field, what steps are being taken to increase openness, and what stands in the way?
- What could help to increase openness?
- What stands in the way?
- What is the Benefits to You ?
- Can you find your own previous work, post-publication and/or pre-publication?
- Can you bring your research materials (data, code, results) with you if you change institutions?
- Can you find the work of your collaborators? Of scientists in other fields that you find interesting?
- Have you reached out to others to collaborate with them after finding interesting results?
- Are people in your field giving and getting credit for work done?

Activity: Think about Open Science (10 min)

(Groups of 3 or 4)

Data Schools

In this activity reflect on your answers to the questions and then compare your thoughts with others in the class

- In your field, what steps are being taken to increase openness, and what stands in the way?
- What could help to increase openness?
- What stands in the way?
- What is the Benefits to You ?
- Can you find your own previous work, post-publication and/or pre-publication?
- Can you bring your research materials (data, code, results) with you if you change institutions?
- Can you find the work of your collaborators? Of scientists in other fields that you find interesting?
- Have you reached out to others to collaborate with them after finding interesting results?
- Are people in your field giving and getting credit for work done?

Persistent Identifiers (PIDs)

A digital persistent identifier (or "PID") is a "long-lasting reference to a digital resource" that is machine-readable and uniquely points to a digital entity.

A persistent identifier used to cite **data**, **software**, journal **articles**, and other types of media *(including presentation slides, blog posts, videos, logos, etc.)*.	An "Open Researcher and Contributor Identifier" (ORCID) provides valid information about a **person**.	The Research Organization Registry (ROR) is a global, community-led registry of open persistent identifiers for **research organizations**
https://www.doi.org/the-identifier/what-is-a-doi/	https://orcid.org/	https://ror.org/about/

Activity: Explore Zenodo and Sign Up! (15 min)

One of the leading ***generalist*** repositories at the moment is Zenodo.

Zenodo is an example of a data repository that allows the upload of research data and creates DOIs.

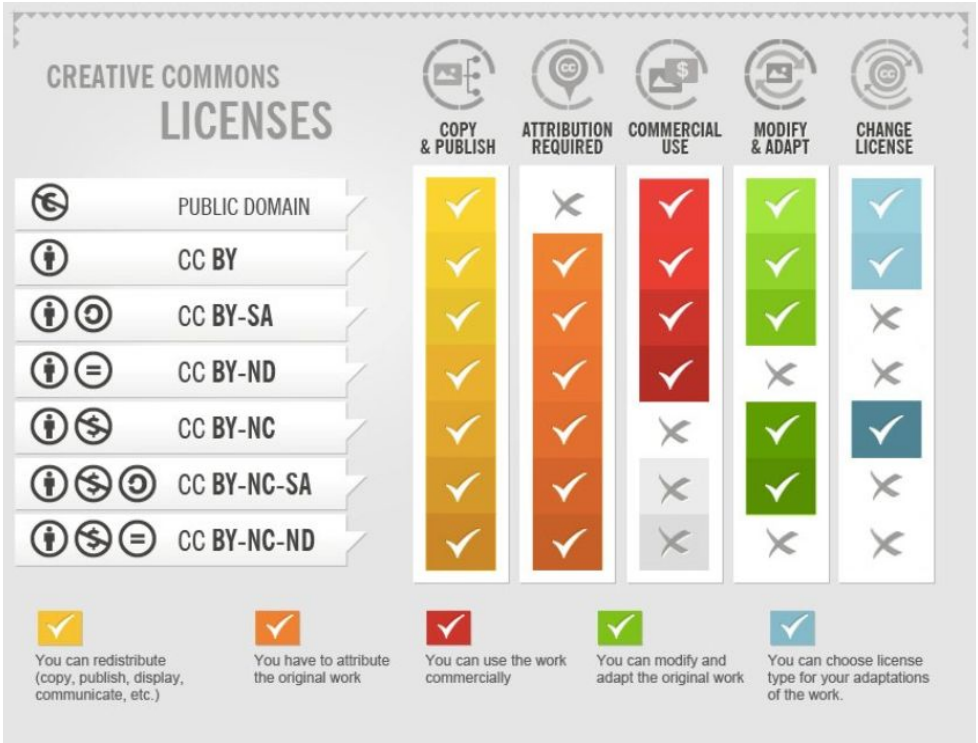
Explore open repositories to familiarize yourself with their structure and available product information.

Optional - Review the following 4.5-minute video (https://www.youtube.com/watch?v=BPVSErzNtME&t=22s) to get an overview of Zenodo and then sign up for an account.

You can use your **ORCID** to sign up if you have one or made one in a previous lesson.

Licenses for Research Data

- Provide information to any potential data **reuser** of their rights
- Ensures clarity
- Creative commons (CC) licensing most used in research



Part of How To Attribute Creative Commons Photos by Foter, licensed CC BY SA 3.0

More about Licenses

- **Licenses can be applied to data, code, and reports, or publications**, and almost any other “creative” output. There are also several different types of licenses and also the case where no license need to apply:

- **Permissive Licenses allow users a wide range of rights including the ability to use, modify, and distribute the work with no restrictions or very few**. Examples of permissive license would be open source software license such as Apache 2.0 or MIT license or the Creative Commons licenses such as Creative Commons Attribution (CC-BY)(https://creativecommons.org/licenses/by/4.0/). Protective Licenses are a legal technique of granting certain freedoms over copies of copyrighted works while including some limitations. This may include copyleft licenses, commercial licenses, or other restrictions.

- **Public Domain is not a license**, but it is an indication that there are no reuse restrictions on the work. Creative Common Zero(https://creativecommons.org/licenses/by/4.0/) is a worldwide public domain mark that indicates that the material is free to use without any restrictions.

Review

Proper Data Management can:

- Increase **value**
- Increase **reproducibility**
- Increase **provenance**
- Increase **integrity**
- Increase **accountability**
- Reduce **risks**
- Reduce **costs**
- Reduce **fraud**

Steps for Planning Outputs in Advance

Data Schools

1. Speaking about it and organizing with your research team
2. Deciding which tools to use
3. Thinking about authorship and credit
4. Engaging with relevant stakeholders and research partners, for example, industry, around open science
5. Identifying repositories for software and data
6. Identifying journals (or other outlets) for publications
7. Highlighting these approaches in your grant and much more

Data Sharing Practice for Zenodo

Data Schools

TL, DR

- Where: https://zenodo.org
- License? CC BY

1. Pick a license (https://creativecommons.org/choose)
2. Drag/drop your files to Zenodo (https://zenodo.org) (or FigShare)

   (The citation file is automatically generated by the data repository)

3. Include description of the data

4. Adopt popular file formats
5. Adopt metadata conventions in your field (e.g., netCDF Climate and Forecast Conventions)

Data Sharing Practice for GitHub

Data Schools

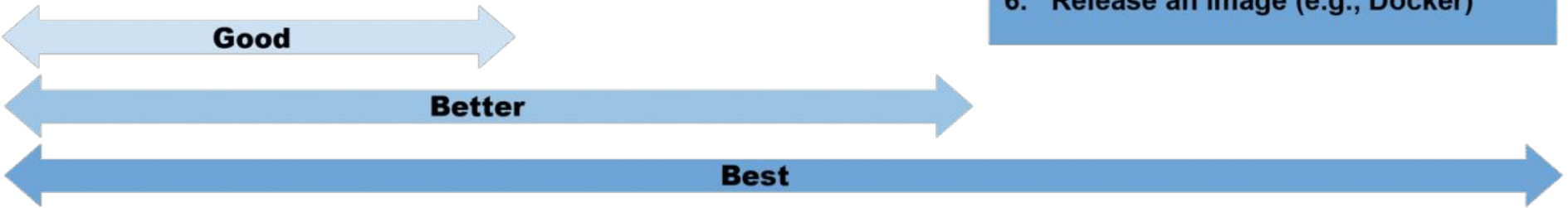
TL, DR

- Where: https://github.com
- License? BSD 3-Claus

1. Pick a license (https://choosealicense.com/)
2. Add your files to GitHub https://github.com (or BitBucket)

3. Include some description (can be comments within code) and basic installation instructions

4. Describe software dependencies (Linux: apt-get, Mac: brew, Windows: Chocolatey; Python: pip)
5. Use continuous integration (e.g., GitHub Actions),
6. Release an image (e.g., Docker)



Activity

In this activity, you will search for a **DOI** for a data set or piece of software that you use, and you will then use the DOI website to “resolve” the DOI name. By “resolving”, this means that you will be taken to the information about the product designated by that particular DOI.

Try the following DOI Numbers

1. 10.9734/jesbs/2022/v35i530421
2. 10.5194/os-13-551-2017
3. 10.1080/10095020.2020.1854981
4. 10.1038/s41586-018-0651-8
5. 10.5067/TERRA-AQUA/CERES/EBAF-TOA_L3B004.1

Open Data Sharing Process

In general, sharing your open data requires the following steps:

1. Make sure your data can be shared
2. Select or identify a repository to host your data
3. Work with your repository to follow their process and meet their requirements
4. Make sure your data is findable and accessible through the repository and is maintained and archived
5. Request a DOI for your data set so that it is easily citable
6. Choose a data license

Activity: How to Enable Reuse of Code

Part 1: Create a test public GitHub repository.

Part 2: Create an archived repository and affiliated DOI.

Part 1: Create a test public GitHub repository

1. Navigate to the login page for GitHub and login. If you haven't already, create a free user account.
2. Create a new repository with this link
3. Type a short, memorable name for your repository. For example, "os-test".
4. Set the repository visibility 'Public' by selecting this option below the repository description.
5. In the following section 'Initialize this repository with:' select 'Add a README file'.
6. Select any license.
7. Click 'Create repository'.
8. You will be automatically directed to your new repository webpage.
9. Get a DOI from the Zenodo application. **We will use:**

    a. https://sandbox.zenodo.org/ to do this.

    b. This offers all the same capabilities as https://zenodo.org but is a testing site! Create a free account if you have not already.

Part 2: Create an archived repository and affiliated DOI

1. Navigate to the Zenodo GitHub page Click on the button 'Connect' to allow Zenodo to access your GitHub repositories.
2. Review the information about access permissions, then click ‘Authorize Zenodo’.
3. Sync your GitHub with Zenodo by clicking 'Sync now' in the upper right corner.
4. To the right of the name of the repository you want to archive (‘os-test’), toggle the button to On.
5. Click on the name of the repository.
6. Click the big green button that has ‘username/os-test’
7. Add a tag ‘test’. You may have to create a new tag for ‘test’ if prompted.
8. Scroll down and click the green ‘publish release’ button
9. Navigate to the Zenodo GitHub page and see the DOI for ‘os-test’
10. Share your DOI

Zenodo archives your repository and issues a new DOI each time you create a new GitHub release . Follow the steps at "Managing releases in a repository to create a new one.

Thoughts on RDM

Sharing data is very good thing indeed….

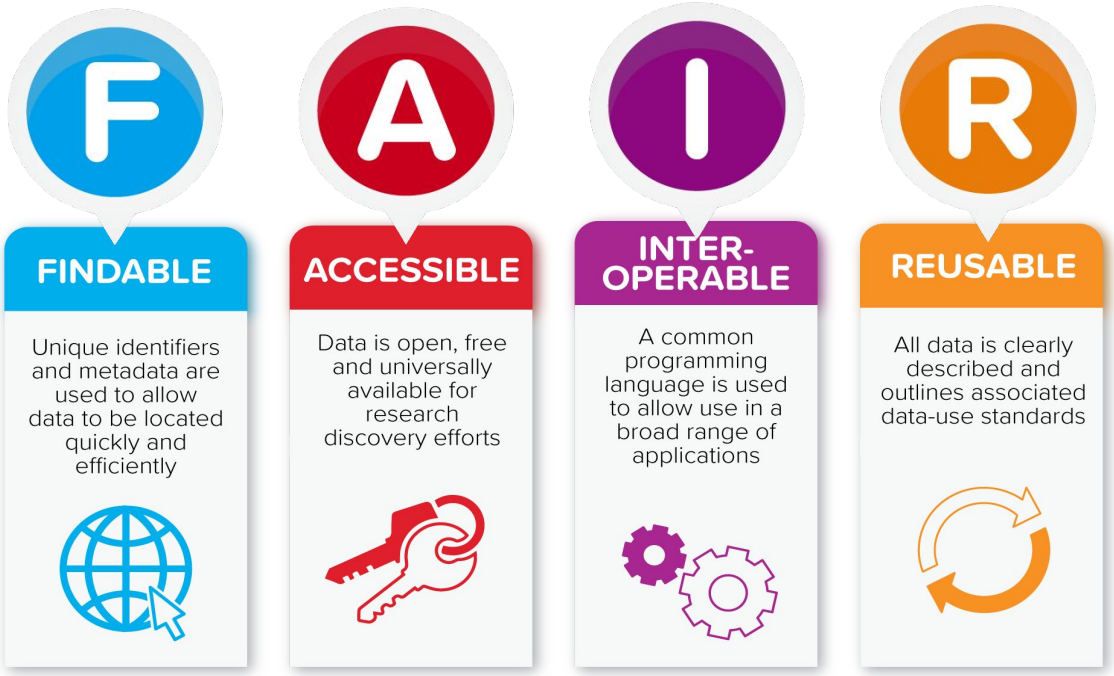
BUT - you need to make sure that the data conforms to good data management practice

Should consider if the data (all of it? some of it?) *should* be open

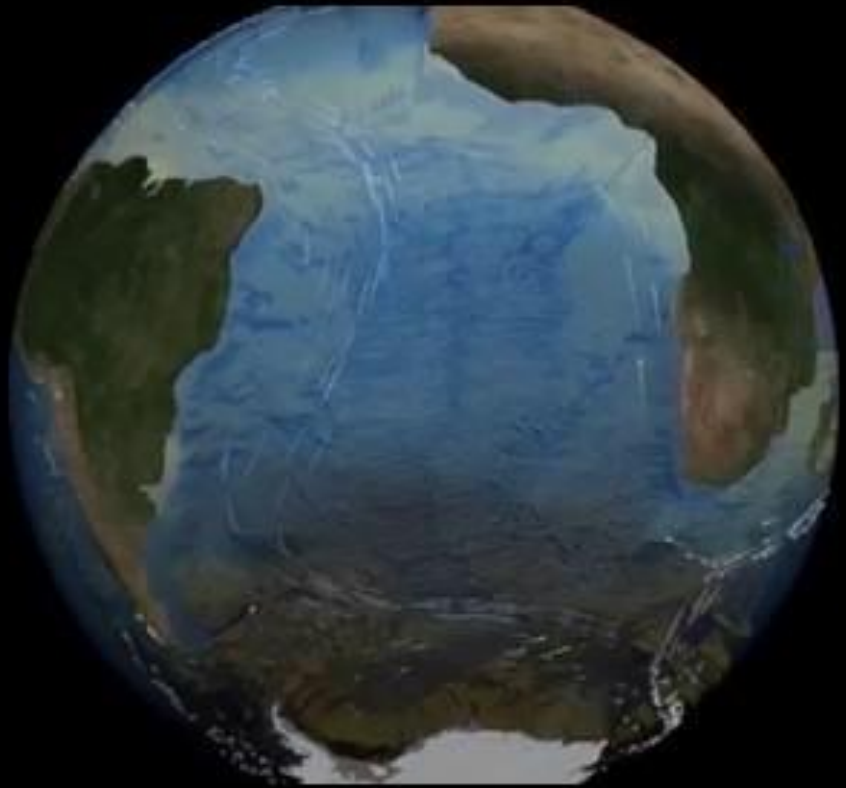
If you want to share data then it should conform to set of principles…..

FAIR

- **F**indable, **A**ccessible, **I**nteroperable & **R**eusable
- Part of a growing movement to increase the value of data by ensuring their long term preservation
- A distillation of good RDM practices

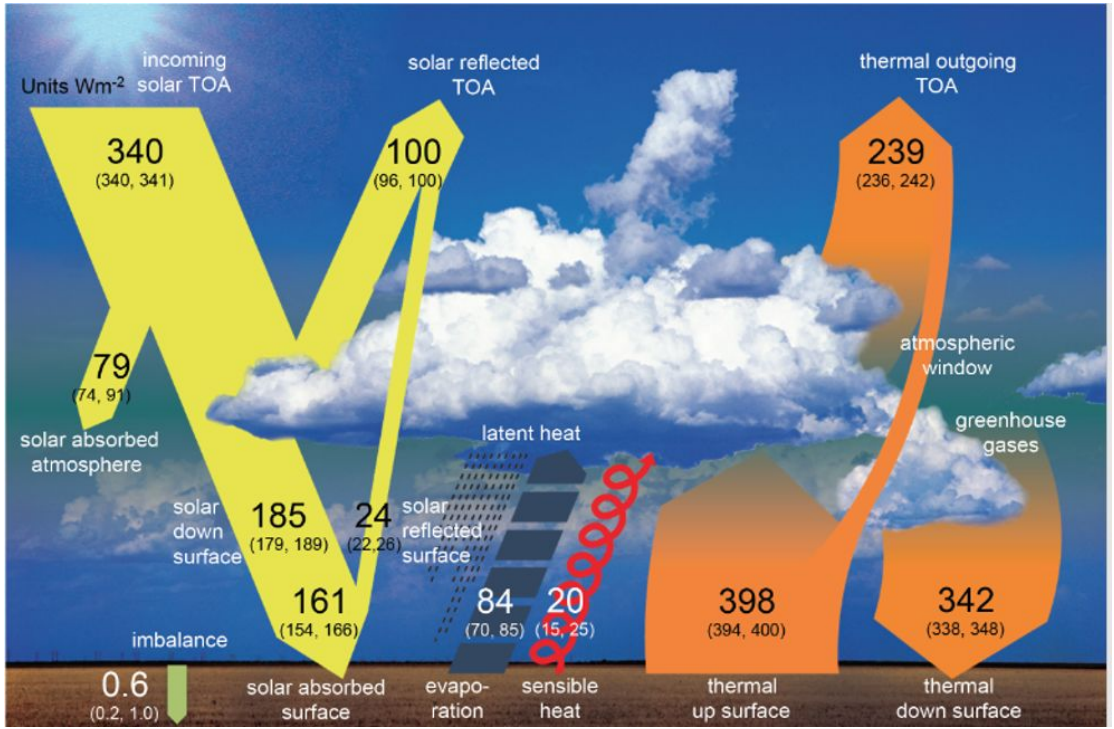


Ocean Observations 101




1.332 x 10^21 Liters of Water
~352 Quintillion Gallons

The need for global ocean observations: Global Ocean Heat Content

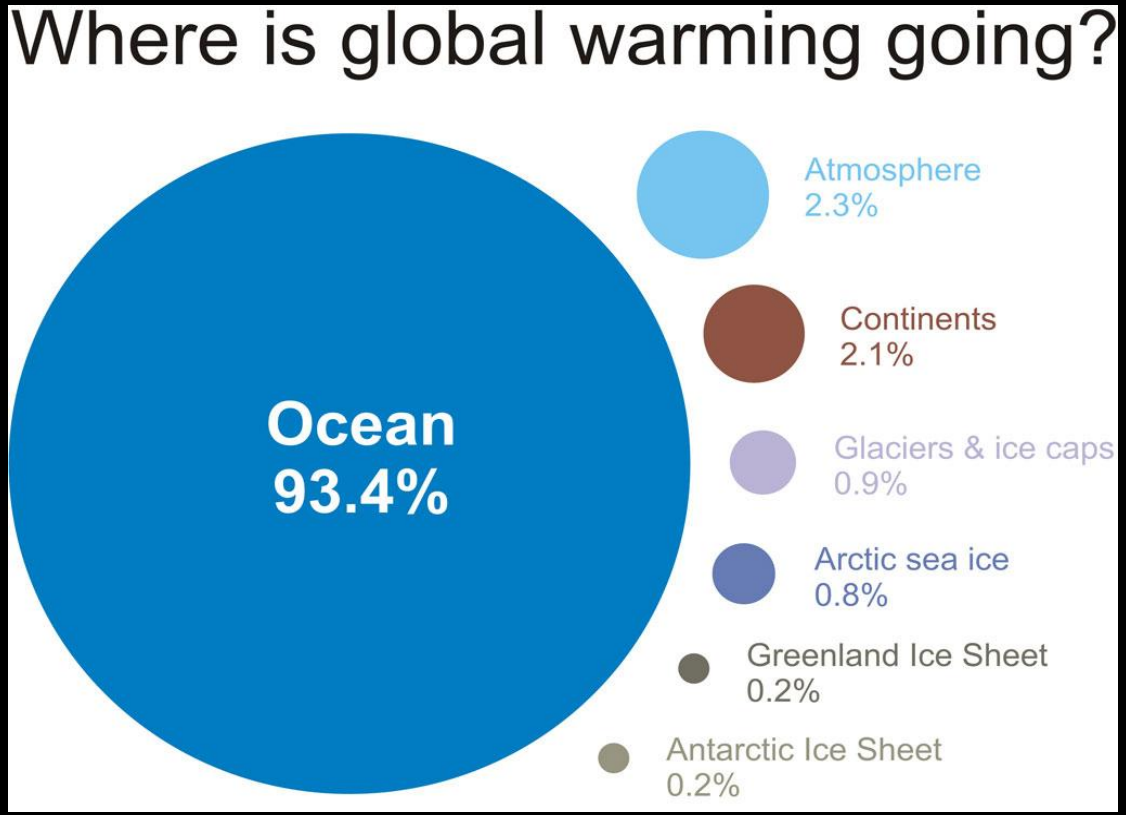


Earth's heat balance
(Source: IPCC 5th Assessment Report)

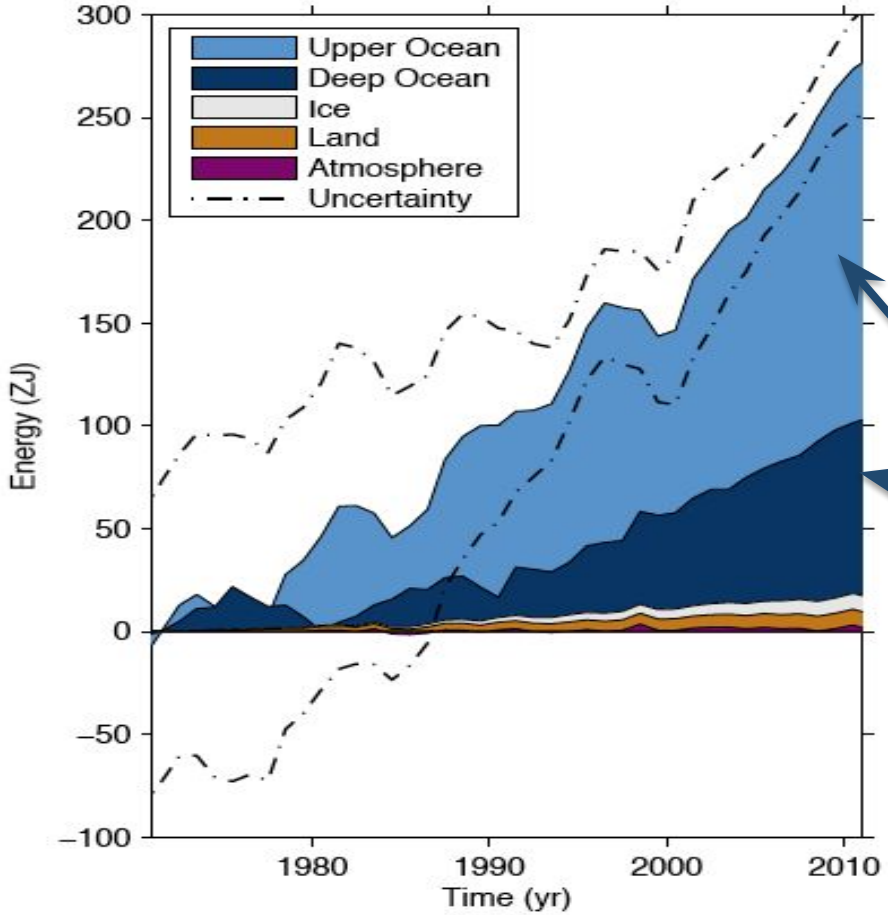
Energy fluxes at the top of the atmosphere are accurately measured by satellites, but the net flux is a small difference of large numbers and is challenging.

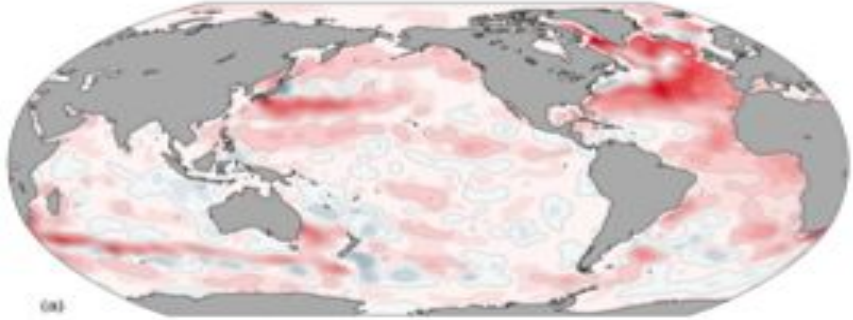
This net heat flux is more accurately measured from temperature changes of the ocean, atmosphere, and cryosphere.

Oceans are the “flywheel of climate”



S. Levitus, J. I. Antonov, T. P. Boyer,O. K. Baranova,H. E. Garcia,R. A. Locarnini,A. V. Mishonov,J. R. Reagan,D. Seidov,E. S. Yarosh, and M. M. Zweng | published 17 May 2012



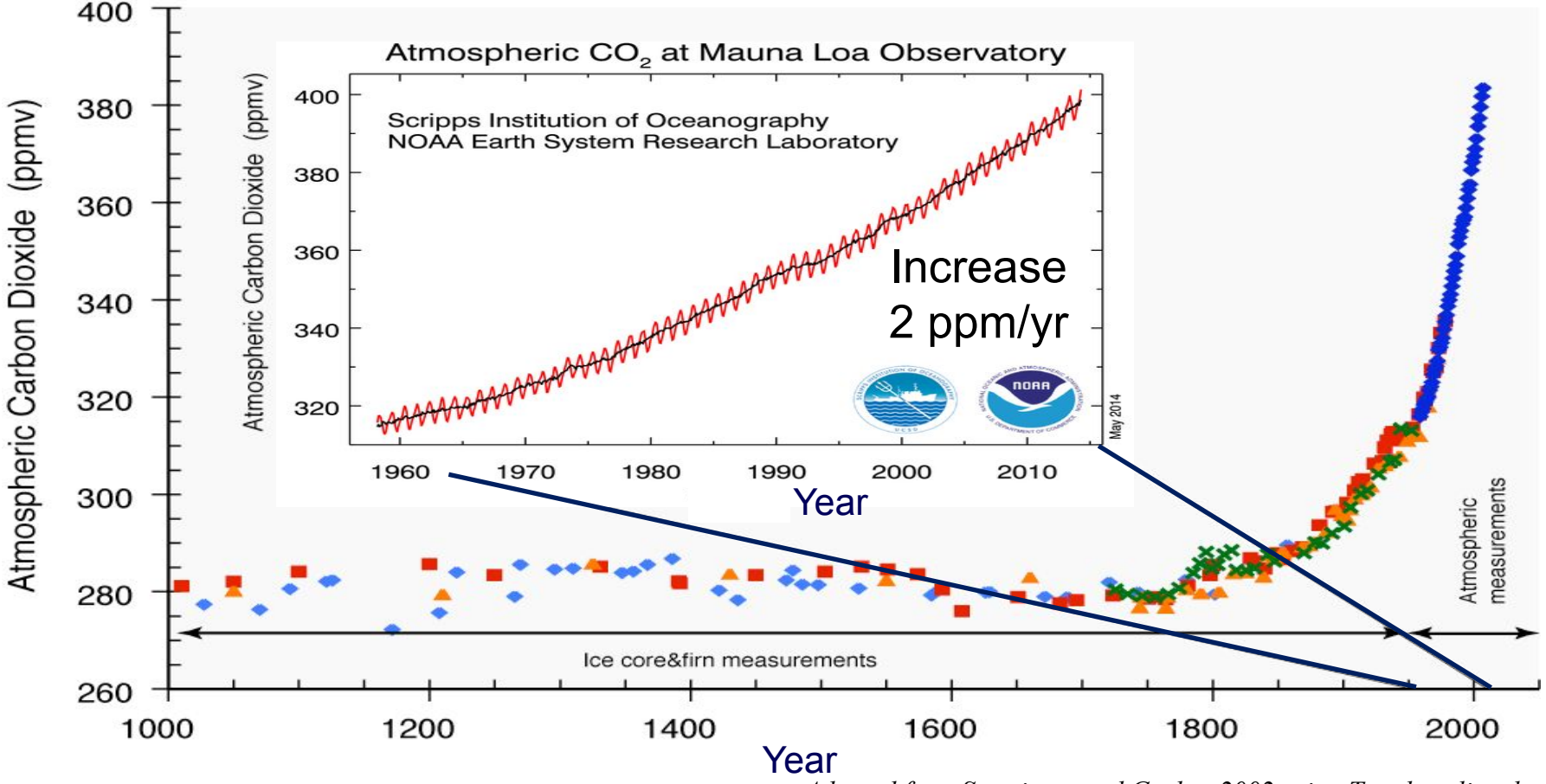


The global energy excess is mostly in the ocean

If the excess heat in the system that is now in the ocean were in the atmosphere, surface air would be 100°C warmer.

Atmospheric CO_2 was steady for at least 1,000 years

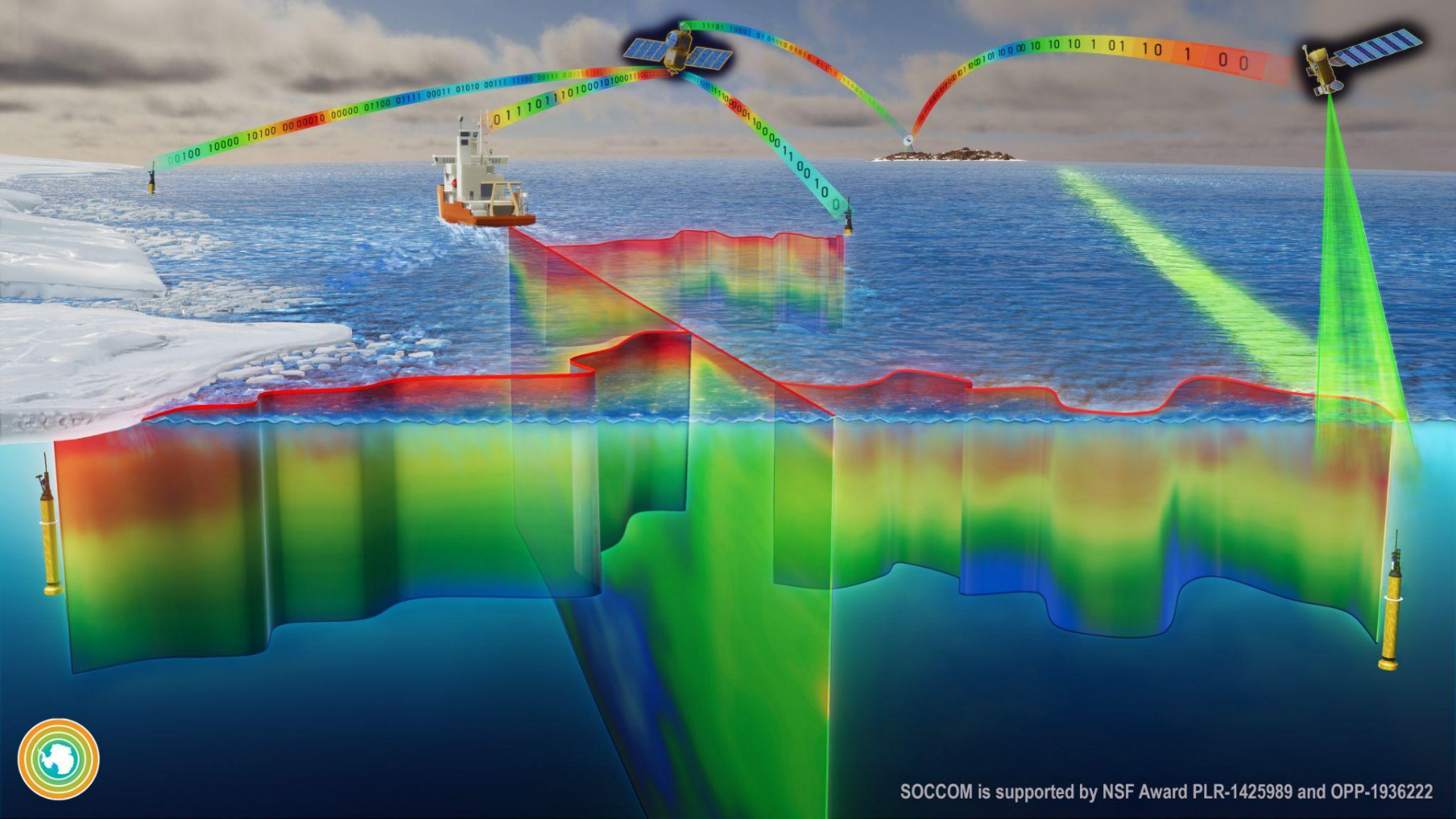
pre-industrial revolution



Adapted from Sarmiento and Gruber 2002 using Trends online data

Ocean Acidification

- Approximately 28% of the CO_2 generated by human activities since the mid-1700s has been absorbed by the oceans.
- Ocean acidity has increased 30% since the start of the industrial age.
- Ocean acidity is projected to increase 100-150% percent by 2100.
- Current rate of acidification is nearly 10x faster than any period over the past 50 million years.



SOCCOM is supported by NSF Award PLR-1425989 and OPP-1936222

Extra Terrestrial Biology

Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 332, NO. 6034 > A BACTERIUM THAT CAN GROW BY USING ARSENIC INSTEAD OF PHOSPHORUS

RESEARCH ARTICLE

A Bacterium That Can Grow by Using Arsenic Instead of Phosphorus

FELISA WOLFE-SIMON, JODI SWITZER BLUM, THOMAS R. KULP, GWYNETH W. GORDON, SHELLEY E. HOEFT, JENNIFER PETT-RIDGE, JOHN F. STOLZ, SAMUEL M. WEBB, PETER K. WEBER, [...], AND RONALD S. OREMLAND +2 authors Authors Info & Affiliations

SCIENCE - 2 Dec 2010 - Vol 332, Issue 6034 - pp. 1163-1166 - DOI: 10.1126/science.1197258

29,019 348

Abstract

Life is mostly composed of the elements carbon, hydrogen, nitrogen, oxygen, sulfur, and phosphorus. Although these six elements make up nucleic acids, proteins, and lipids and thus the bulk of living matter, it is theoretically possible that some other elements in the periodic table could serve the same functions. Here, we describe a bacterium, strain GFAJ-1 of the Halomonadaceae, isolated from Mono Lake, California, that is able to substitute arsenic for phosphorus to sustain its growth. Our data show evidence for arsenate in macromolecules that normally contain phosphate, most notably nucleic acids and proteins. Exchange of one of the major bio-elements may have profound evolutionary and geochemical importance.

Research published in 2010 suggested that the bacterium GFAJ-1 could use arsenic instead of phosphorus in its essential biomolecules, like DNA. This was a groundbreaking idea as phosphorus is considered a fundamental building block of life as we know it.

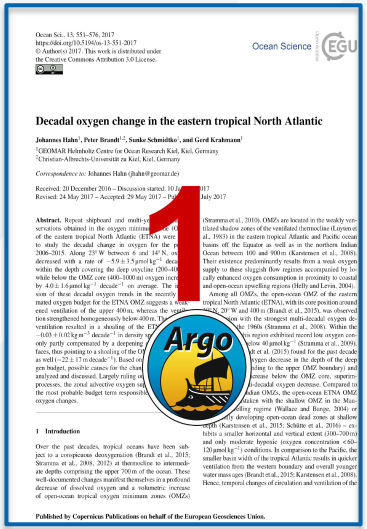
Significance: If true, the claim would have major implications for the search for extraterrestrial life. It would expand our understanding of the conditions under which life can exist and potentially open up new avenues for exploration in environments previously thought to be inhospitable.

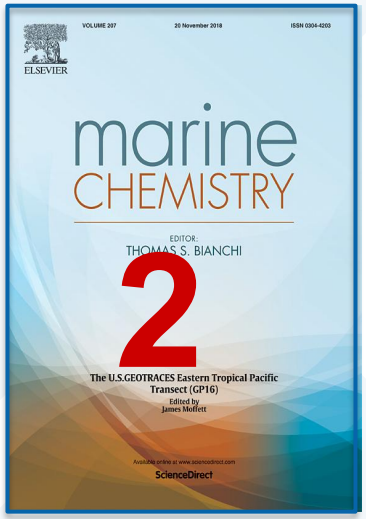
Workshop Time

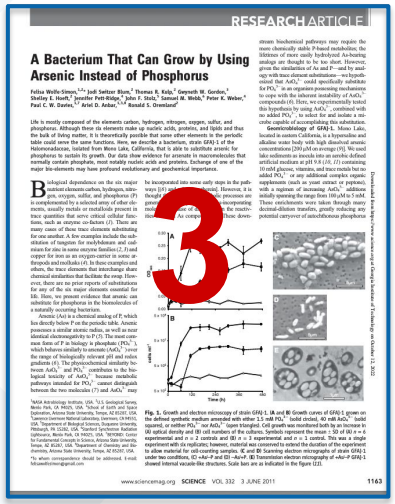
Problem Description: Teamwork!

Your interdisciplinary science team will quickly review one of these peer-reviewed papers

https://bit.ly/2023_RDM_papers







Problem Description

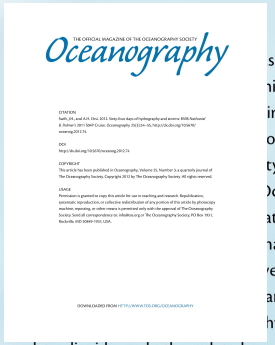
- Your group has decided to try to replicate the results from the research described in the paper
- Your assignment is as follows:
  1. Scan the paper, look for clues as to where the data used in this research might reside
  2. Find and *try* to download those data

  3. **Be mindful throughout of what makes the data easy/hard to:**
     - Locate
     - Download
     - READ/UNDERSTAND (possibly with other data)

Example

Sixty-Four Days of Hydrography and Storms: RVIB *Nathaniel B. Palmer's* 2011 S04P Cruise

BY JAMES H. SWIFT AND ALEJANDRO H. ORSI



s system-
ic sec-
irculation
ountry's
ty and
Ocean
ate
al Ocean
verarching
anges
hwater,

carbon dioxide, and other related parameters.

In brilliant Antarctic weather and rarely seen open waters in McMurdo Sound, we set out on the icebreaking research vessel *Nathaniel B. Palmer* from the ice pier at the US Antarctic

In addition, we aimed to close off key CLIVAR meridional transects to the Antarctic shelf break, including completion of transects along 150°W and 170°W.

With nominal spacing of 30 nm, each station consisted of a full-depth deployment of a 36-place rosette/CTD equipped with dual temperature/conductivity channels, pressure and dissolved oxygen instruments, a reference thermometer, a transmissometer, a fluorometer, an altimeter, and an acoustic Doppler current profiler (ADCP). Water samples were collected for measurements of salinity, dissolved oxygen, nutrients, chlorofluorocarbons, dissolved inorganic and organic carbon, total alkalinity, pH, colored dissolved organic matter

ADCP, surface temperature/salinity/pCO_2, and other seawater properties, meteorology, solar radiation, and aerosols/precipitation. US data and accompanying documentation are publicly available at the CLIVAR and Carbon Hydrographic Data Office (via http://ushydro.ucsd.edu) and the Carbon Dioxide Information Analysis Center (http://cdiac.ornl.gov).

As we neared Cape Adare to start the first station, winds rose well past 30 knots and continued to roughen the seas during the day. This weather was a taste of the future, because storms frequently interrupted our work (e.g., 105 hours were lost in the first two weeks of the cruise alone), but the new data were fascinating from the start.

ADCP, surface temperature/salinity/pCO_2, and other seawater properties, meteorology, solar radiation, and aerosols/precipitation. US data and accompanying documentation are publicly available at the CLIVAR and Carbon Hydrographic Data Office (via http://ushydro.ucsd.edu) and the Carbon Dioxide Information Analysis Center (http://cdiac.ornl.gov).

Your Workspace

Use the provided template 1-2 slides on what you found, and be prepared to discuss your findings in about 12 minutes. You can decide on a presenter any time between now and when you present.

https://bit.ly/2023_RDM_papers

Template: Your Team's Research Data Assessment Report

- **Our Team consisted of …**
  - Name 1 …. Name N
- **Persistent Identifier for the data**
  - PID
- **We found the data we were looking for here:**
  - Name of repository / data center and URL
- **The data was discoverable/accessible**
  - Were you able to read the data with ODV or other software?
  - Could not find the data (why?)
- **Overall FAIRness assessment (1=*awful* to 10=*great*)**

The Future is Now Browser

How Emerging Technology Like AI is Changing How We Do Science

- Use of AI for Literature review
- The ever-increasing volume of scientific literature has made it challenging for researchers to stay abreast of recent articles and find relevant older ones.
- AI tools can be used to create personalized recommendations for relevant articles as well as create summaries of them in various formats.
- Some examples of these tools include SciSummary, , SummarizeBot, Scholarcy, Paper Digest, Lynx AI, , TLDR This

- **Possible drawbacks when using these tools include:**
  - Potential introduction of biases
  - Insufficient contextual understanding or interpretation
  - Possible inability to handle complex technical language
  - Incorrectly identifying key points

AI tools for Code Generation

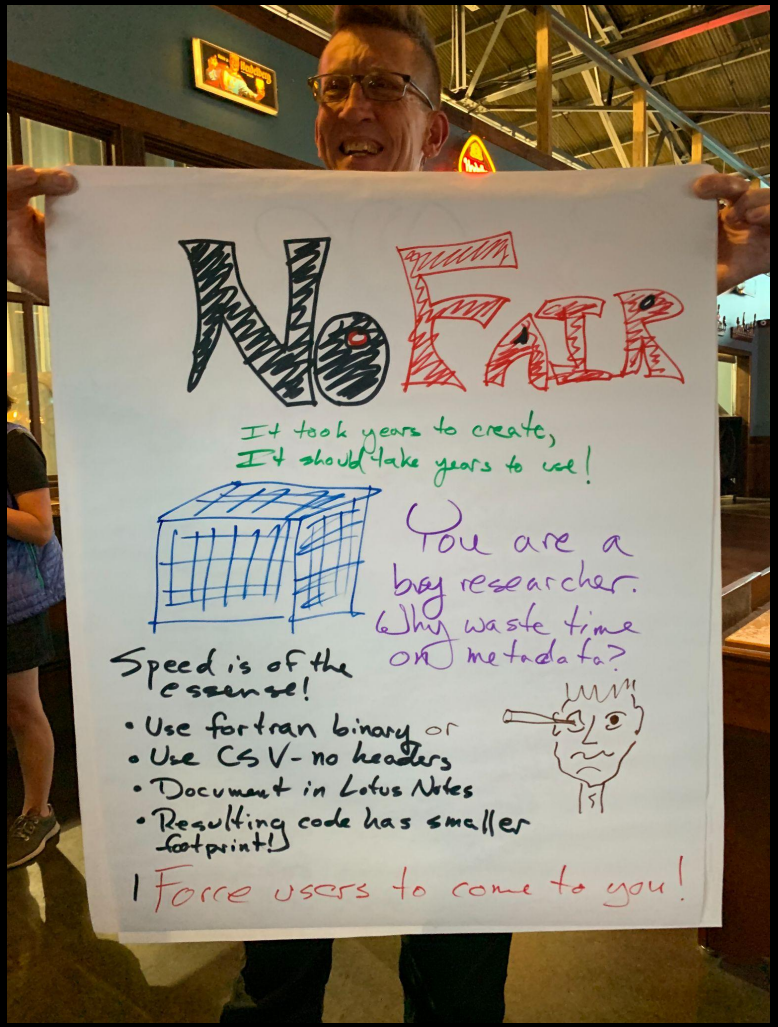
AI tools can be used to generate code to perform analysis tasks and translate between programming languages. Some examples of these tools include Co-Pilot, , Codex, ChatGPT, and AlphaCode, and NotebookLM (Gemini)

Usage tip: Popular large language models can be used to generate code, but it has been noted by many that breaking down tasks and using careful prompts helps generate better results.

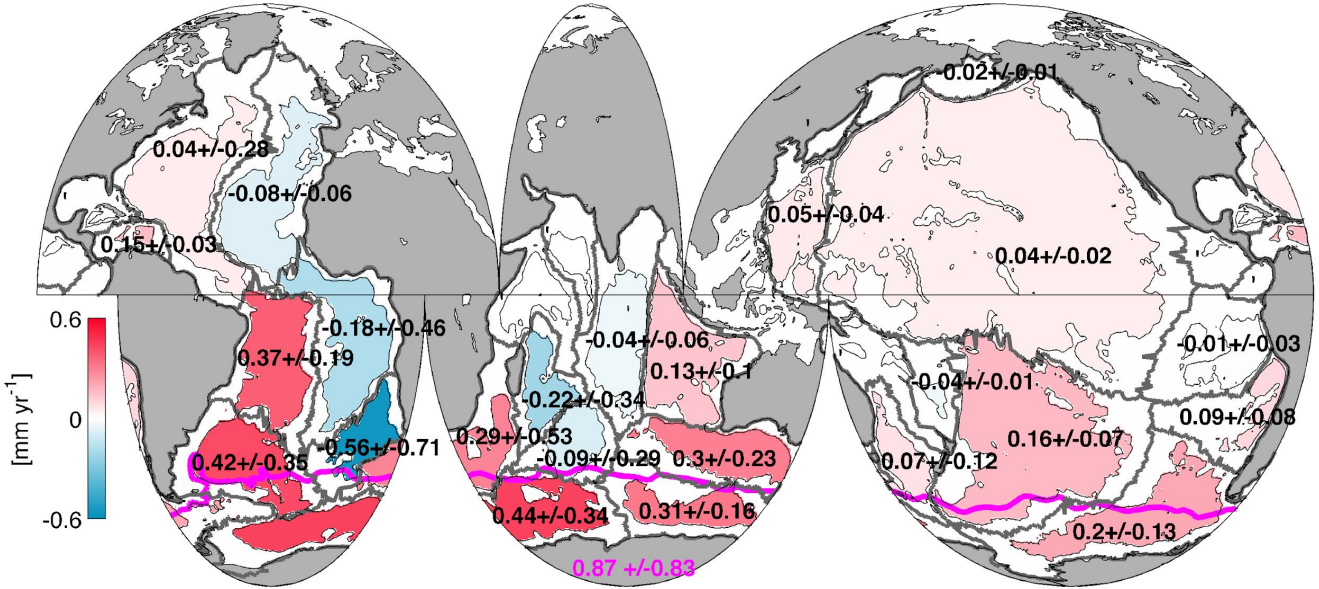
AI tools can be used to generate text, summarize background materials, develop key points, develop images and figures, and conclusions. Using these tools may help non-native speakers communicate science in different languages more clearly. Additionally, they could be helpful to develop plain-language summaries, blog posts, and social media posts.

AI tools for science

- AI tools for science are developing rapidly.
- The science community's understanding of how to ethically and safely use AI is just developing as its use in research expands rapidly.
- The guidelines above offer a snapshot in time and will likely continue to evolve.
- If you choose to use these tools for scientific research, carefully consider how much to rely on them and how their biases may impact results, as cautioned in this Nature article.
- The internet has transformed the world and AI tools are likely to do the same. As with any tool, it is important they are used for the appropriate purpose and in an ethical manner.


No FAIR
It took years to create,
It should take years to use!
You are a
big researcher.
Why waste time
on metadata?
Speed is of the
essense!
• Use fortran binary or
• Use CSV - no headers
• Document in Lotus Notes
• Resulting code has smaller
footprint!
Force users to come to you!

Comparisons and Connections



Bottom Water warming from 1990's to 2000's

Purkey and Johnson (2010)

GO-SHIP CTD data are inputs to the OWC algorithms that are used to estimate the time-varying correction of conductivity measurements from **Argo floats.**

University of California CDL California Digital Library