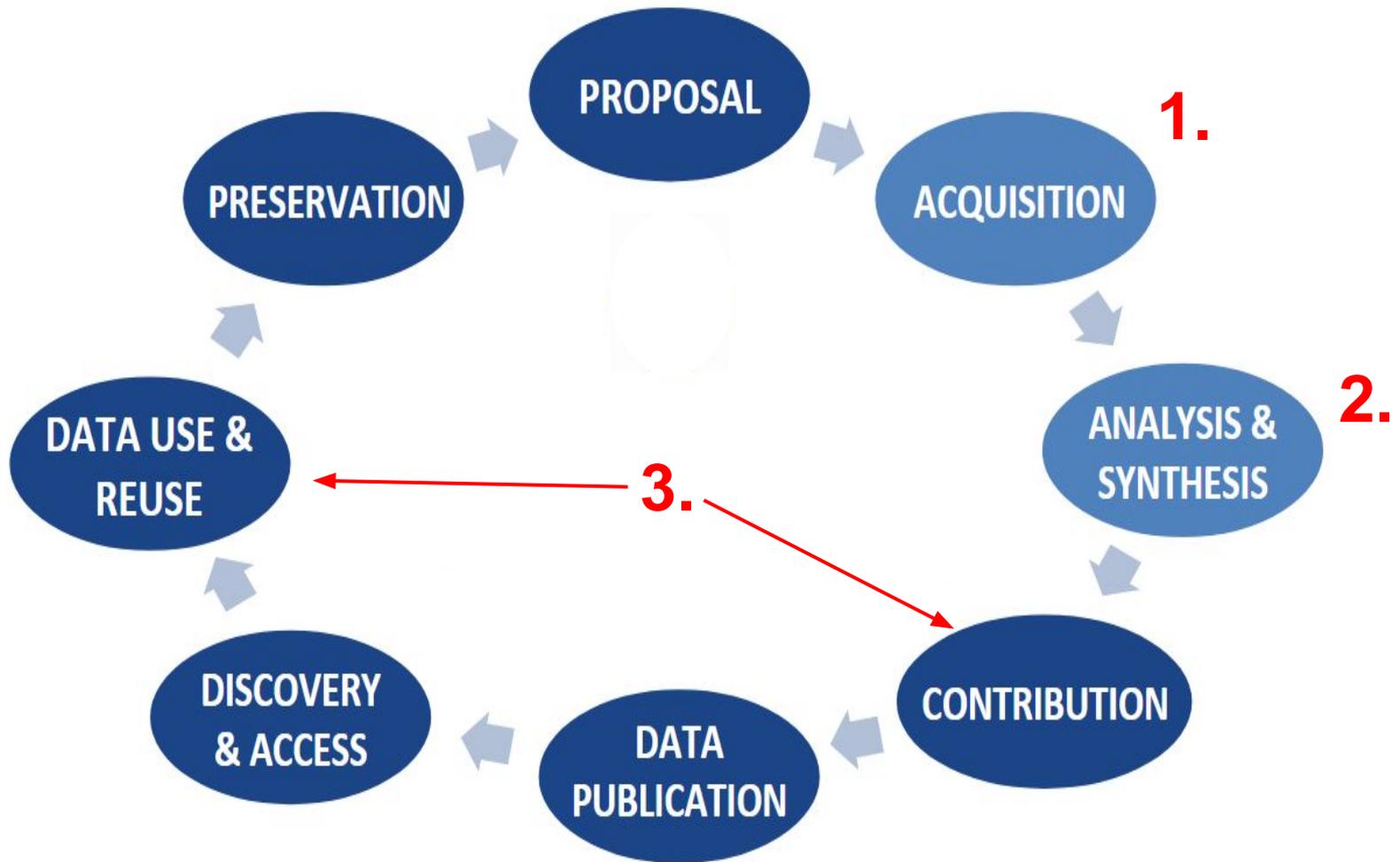


# Navigating an Ocean of Data Training Camp

## **Module 4: Best practices in the ocean sciences**

Stace Beaulieu and Danie Kinkade  
January 23, 2020

# Module 4



THE DATA LIFE CYCLE

# How are we different than the definition provided for “data scientist”?

- Data Scientists develop methods for storing, analyzing, and presenting data

# We are data creators!

Types of data created at WHOI range from observational (sensors and physical samples) to experiments (with observations) to models

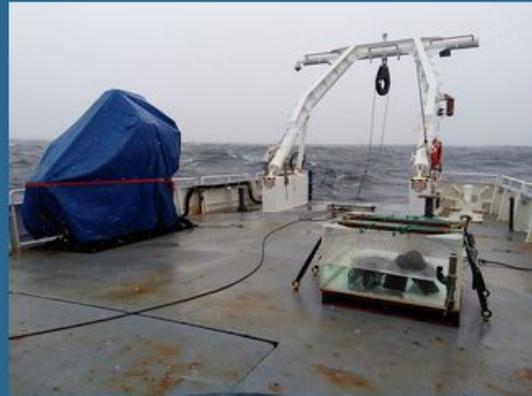


ABOUT RESEARCH DATA EDUCATION



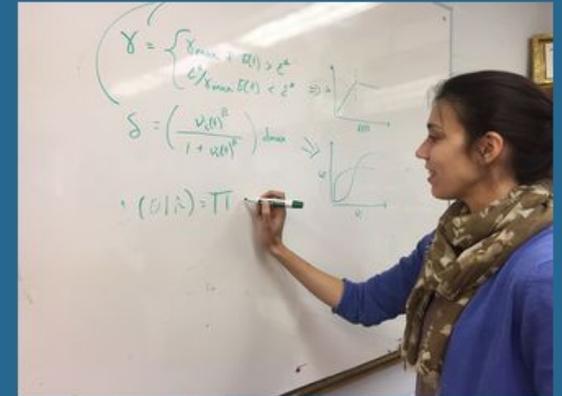
Deploying and recovering the CTD/rosette from R/V Endeavor during the 2018 Winter Transect cruise EN608. Image credit: Jacob Strock, University of Rhode Island.

Observations



Incubation experiments on the deck of R/V Endeavor during the 2018 Winter Transect cruise EN608. Image credit: Susanne Menden-Deuer, University of Rhode Island.

Experiments



The NES-LTER project includes mathematical models of ecological systems. (Photo courtesy Michael Neubert)

Models

Example: NES-LTER project (<https://nes-lter.who.edu/>)

# We are data creators!

<http://www.whoi.edu/data/>

The screenshot shows the top navigation bar of the Woods Hole Oceanographic Institution website. The logo is on the left, and navigation links for 'PRESS ROOM', 'SHOP WHOI', and 'DIRECTORY' are on the right. Below the logo are buttons for 'WHO WE ARE', 'WHAT WE DO', 'KNOW YOUR OCEAN', 'JOIN US', and 'DONATE'. The main content area features a 'WHAT WE DO' sidebar with a 'Data & Repositories' section. The main heading is 'Data & Repositories' with a 'Data and Repositories Based at WHOI' sub-heading. There are social media share icons for Facebook, Twitter, Email, and a plus sign. The content is organized into four sections, each with a title, a small image, and a text description.

**Woods Hole Oceanographic INSTITUTION**

PRESS ROOM SHOP WHOI DIRECTORY

WHO WE ARE WHAT WE DO KNOW YOUR OCEAN JOIN US DONATE

WHAT WE DO

▼ Understand

- Areas of Research
- › Departments, Centers & Labs
- › Programs & Projects
- Scientific Facilities & Services

▼ Data & Repositories

- Other National and International Repositories

› Explore

## Data & Repositories

SHARE THIS: [f](#) [t](#) [e](#) [+](#)

### Data and Repositories Based at WHOI

#### Alvin Frame-Grabber

Contacts: Steve Lerner

The Frame-Grabber imaging system mounted on the submersible Alvin provides Web access to video imagery co-registered with vehicle navigation and attitude data for shipboard analysis, planning deep-submergence research cruises, and review of data following research expeditions.

#### Argo Float Data

View the locations of WHOI's floats being used as part of the ARGO project to observe the ocean and to predict climate change. See their launch positions, track the paths they have followed, and pinpoint the locations of

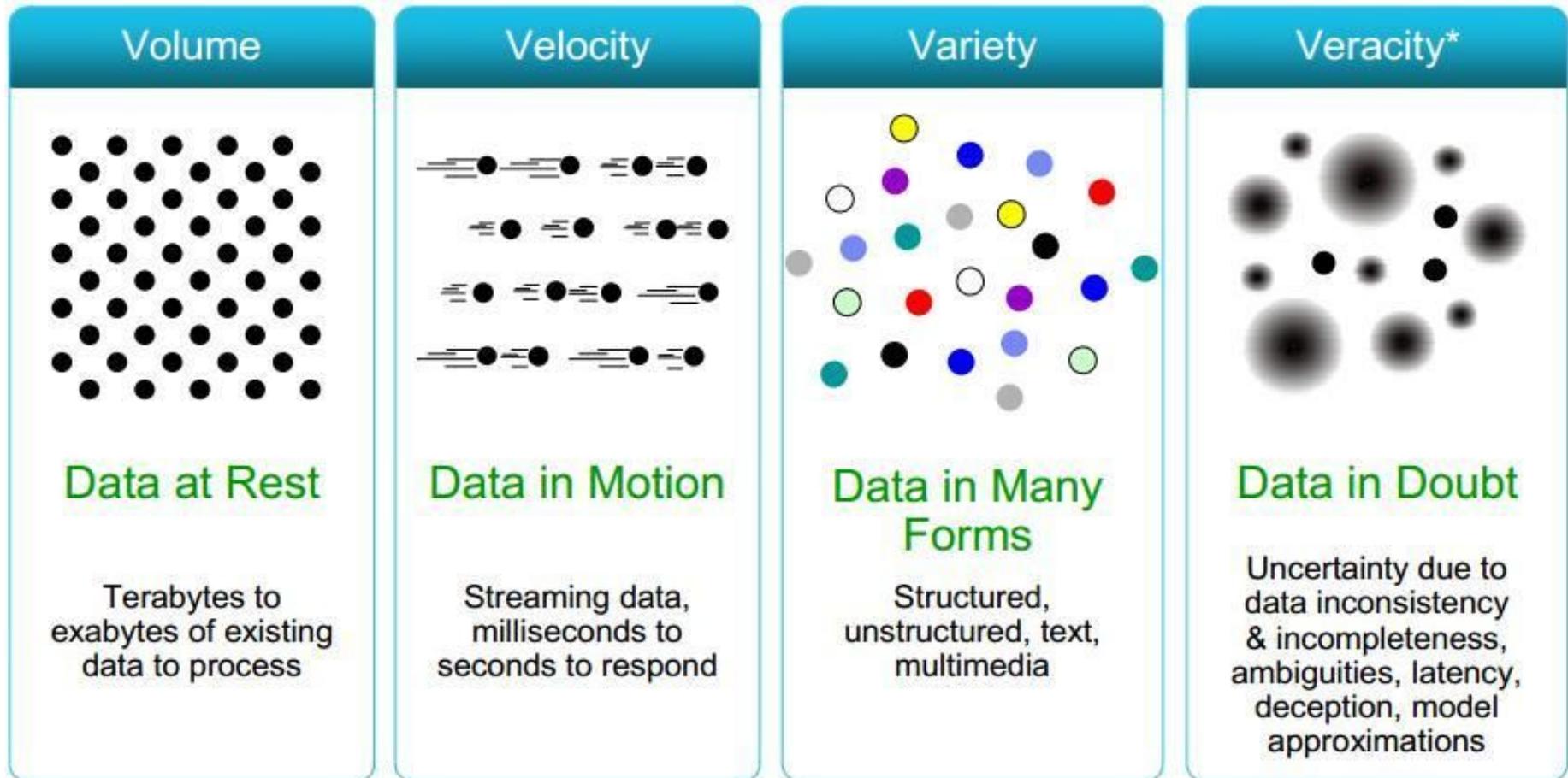
#### Seafloor Sediments Data Collection

The collection and analysis efforts of WHOI's Seafloor Sediments Laboratory are available for browsing. A collection of more than 14,000 archived marine geological samples recovered from the seafloor. The inventory includes long, stratified sediment cores, as well as rock dredges, surface grabs, and samples collected by the submersible Alvin.

#### Upper Ocean Processes Mooring Data

Data collected during the past 25 years by UOP surface moorings that measure meteorological conditions and physical properties of the upper ocean. Includes data gathered along the coasts and the open ocean during various weather conditions and from a range of geographic regions, from subarctic regions to the tropics.

# You experience Big Data when your ability to manage the data exceeds any of these V's:



# We create a variety of high volume and high velocity data

**Module 3**  
**highlighted**  
**high volume data**

e.g., near-real-time observations

Example: <https://oceanobservatories.org/data/>

The screenshot shows the OOI website interface. At the top right, there are links for 'Mailing List', 'OOI Knowledge Base', 'Glossary', and 'Helpdesk'. A search bar is located in the top right corner. The main header features the OOI logo and a banner image of a yellow autonomous underwater vehicle (AUV) with solar panels. Text on the banner states 'The OOI is funded by the National Science Foundation' and includes the NSF logo. Below the banner is a yellow navigation bar with dropdown menus for 'OOI Data', 'The Observatory', 'Community', 'Researchers', 'Educators', 'Science', 'Events & Updates', and 'About'. The main content area is titled 'Data Products' and is divided into three columns: 'Air-Sea Interface', 'Seafloor/Crust', and 'Water Column'. Each column lists various data products with their respective acronyms.

**OOI Data** ▾ **The Observatory** ▾ **Community** ▾ **Researchers** ▾ **Educators** ▾ **Science** ▾ **Events & Updates** ▾ **About** ▾

## Data Products

### Air-Sea Interface

- Air Temperature (TEMPAIR)
- Air Temperature at 2 m (TEMPA2M)
- Barometric Pressure (BARPRES)
- CO2 Mole Fraction in Atmosphere (XCO2ATM)
- CO2 Mole Fraction in Surface Sea Water (XCO2SSW)
- Direct Oceanic Flux of Heat

### Seafloor/Crust

- 16s rRNA sequence of filtered physical sample (DNASAMP)
- Benthic Flow Rates (BENTHFL)
- Broadband Acoustic pressure waves (HYDAPBB)
- Broadband Frequency (HYDFRBB)
- Broadband Ground Acceleration (GRNDACC)
- Broadband Ground Velocity

### Water Column

- Bottom Pressure (IESPRES)
- Conductivity (CONDWAT)
- Density (DENSITY)
- Downwelling Spectral Irradiance (SPECTIR)
- Echo Intensity (ECHOINT)
- Fluorometric CDOM Concentration (CDOMFLO)
- Fluorescence Chlorophyll Concentration

# We have best practices for creating these data

“A community best practice is a methodology that has repeatedly produced superior results relative to other methodologies with the same objective.”

<https://doi.org/10.1029/2018EO096533>



## What's the Best Way to Responsibly Collect Ocean Data?

Evolving and Sustaining Oceans Best Practices Workshop; Paris, France, 15–17 November 2017



Crew aboard the R/V *Atlantic Explorer* deploy equipment that will collect data for the Bermuda Atlantic Time Series. Credit: Juliet Hermes

# <https://www.oceanbestpractices.org/>



All Fields ▾ Search OceanBestPractices Advanced ▾ 🔍 ✕

Search

[Search Tips](#)

Search **926 documents** tagged with **121029 terms** from **6 terminologies**

# We may use physical samples to create these data

It's not just data that need an identifier...  
it's also the physical samples from which the data derive!

Photo by Meg Tivey



Email from Meg when we sent  
out last year's workshop  
announcement:

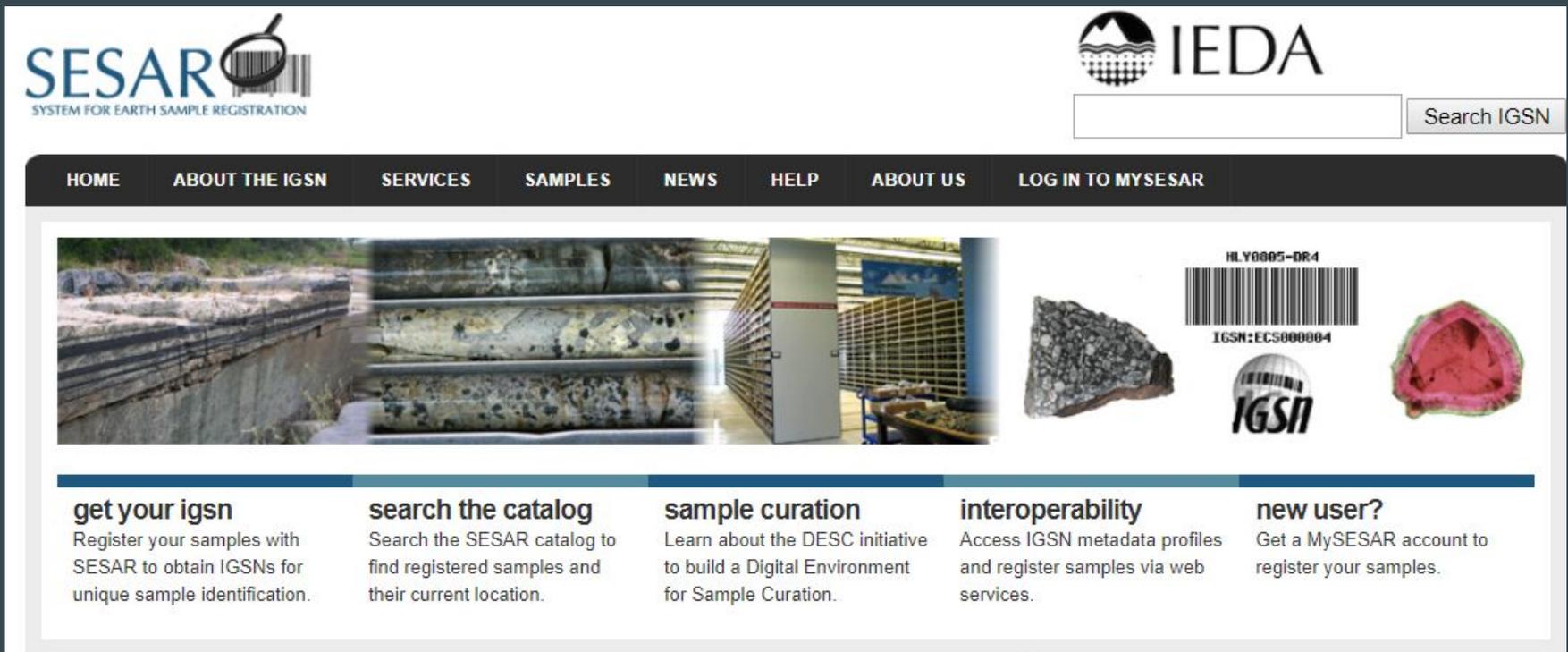
“Perfect timing (for me). I am  
filling out my IGSN metadata  
registration excel  
spreadsheets at this moment!”



# International Geo Sample Numbers (IGSNs)

The IGSN is a persistent unique identifier for physical samples and specimens that eliminates the problems associated with the ambiguous naming of samples. In the U.S., you can obtain IGSNs using the System for Earth Sample Registration (SESAR) at IEDA Data Facility.

See <https://igsn.github.io/overview/> and <http://www.geosamples.org/>



**SESAR**  
SYSTEM FOR EARTH SAMPLE REGISTRATION

**IEDA**

Search IGSN

HOME ABOUT THE IGSN SERVICES SAMPLES NEWS HELP ABOUT US LOG IN TO MYSESAR

**get your igsn**  
Register your samples with SESAR to obtain IGSNs for unique sample identification.

**search the catalog**  
Search the SESAR catalog to find registered samples and their current location.

**sample curation**  
Learn about the DESC initiative to build a Digital Environment for Sample Curation.

**interoperability**  
Access IGSN metadata profiles and register samples via web services.

**new user?**  
Get a MySESAR account to register your samples.

# Let's look for Meg's recently registered International Geo Sample Numbers (IGSNs)

In U.S.: System for Earth Sample Registration (SESAR)  
at IEDA Data Facility

***Take the next 5 min:***

- Search for Meg's samples in SESAR:  
<http://www.geosamples.org/catalogsearch>
- What kinds of information can you find out about her samples?

Hint: Advanced Settings > Registrant

# What kinds of information can you find out about her samples?

Advanced Settings

[Clear](#)

Registrant: Margaret Tivey (mktivey@whoi.edu)

Total sample counts are listed, but only public sample metadata can be accessed. Please note that downloading large datasets may take some time.

20 Grab(s)

14 Core(s)

Page 1 of 1

[Download all public samples](#) of 20 Grab(s)

IGSN	Sample Name	Object Type	Material:Classification	Location
<a href="#">IEMKT000F</a>	J2-1101-1-R1	Grab	Rock:Hydrothermal>Sulfide	Endeavour Segment: near Main Endeavour Field
<a href="#">IEMKT000G</a>	J2-1101-2-R1	Grab	Rock:Hydrothermal>Sulfide	Endeavour Segment: near Main Endeavour Field
<a href="#">IEMKT000H</a>	J2-1101-2-R2	Grab	Rock:Igneous>Volcanic	Endeavour Segment: near Main Endeavour Field
<a href="#">IEMKT000I</a>	J2-1101-3-R1	Grab	Rock:Hydrothermal>Sulfide	Endeavour Segment: near Main Endeavour Field
<a href="#">IEMKT000J</a>	J2-1101-4-R1	Grab	Rock:Hydrothermal>Sulfide	Endeavour Segment: near Main Endeavour Field
<a href="#">IEMKT000K</a>	J2-1101-5-R1	Grab	Rock:Hydrothermal>Sulfide	Endeavour Segment: near Main Endeavour Field

IGSN: IEMKT000I



**IGSN:** IEMKT000I  
**Sample Name:** J2-1101-3-R1  
**Other Name(s):**  
**Sample Type:** Grab  
**Parent IGSN:** Not Provided

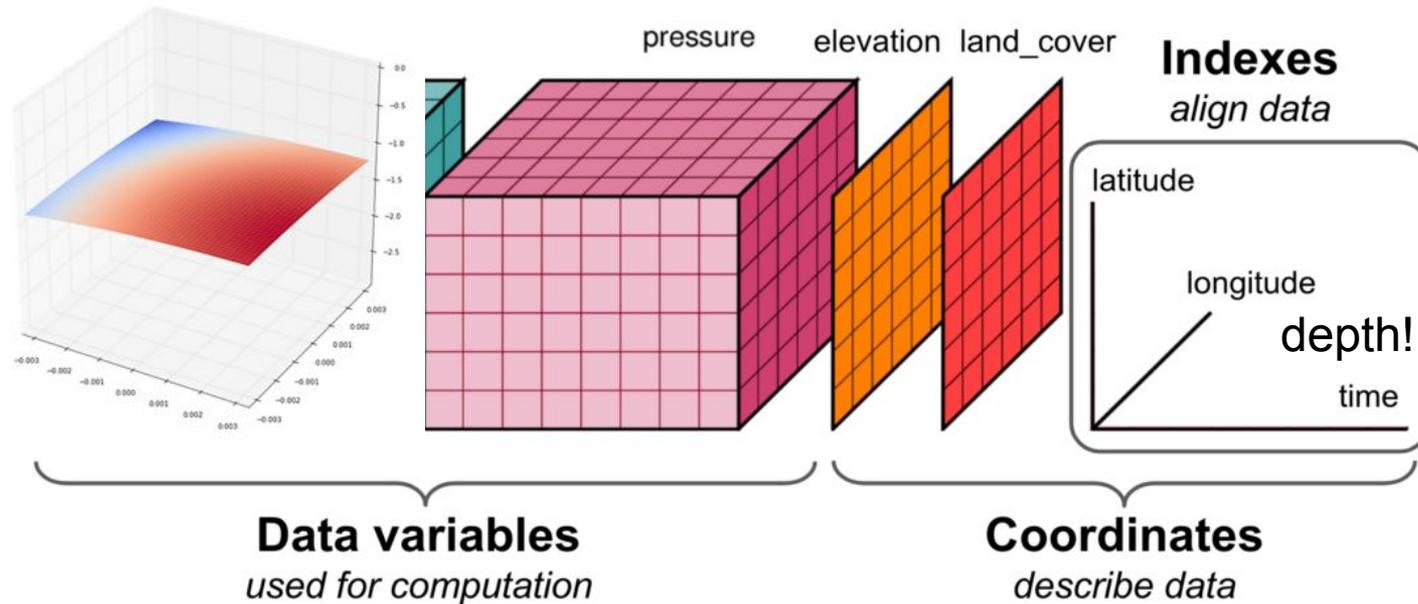
**Description**

Material: Rock

Latitude (WGS84): 47.9552676  
Longitude (WGS84): -129.0960122  
Elevation Start: -2160.73 meters  
Field Program/Cruise: KM1812  
Platform Type: Ship  
Platform Name: Kilo Moana  
Launch Type: ROV  
Launch Platform Name: Jason II  
Launch ID: J2-1101  
Collection Start Date: Not Provided

# Importance of the 4 D's in our domain

dimensions



From Module 1, definition Earth Science Data Analytics:

"The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data, encompassing varieties of data types..."

<https://medium.com/pangeo/step-by-step-guide-to-building-a-big-data-portal-e262af1c2977>

(blog post by MIT alum Ryan Abernathy)

# Controlled vocabularies for (units and) variables

NERC Vocabulary Server

Poll results from  
yesterday:

<https://www.bodc.ac.uk/resources/vocabularies/>

Have you heard of a  
“controlled vocabulary”?

- Yes = 0?

**F and I in FAIR data**

↑ -- **Sigma-theta of the water body by thermosalinograph and computation from salinity and potential temperature using UNESCO algorithm --**

URI	<a href="http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/">http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/</a>
Identifier ()	SDN:P01::SIGTSG01
Preferred label (en)	<b>Sigma-theta of the water body by thermosalinograph and computation from salinity and potential temperature using UNESCO algorithm</b>
Alternative label (en)	STTSG
Definition (en)	Computed by UNESCO SVAN function using potential temperature
Version Info ()	1
Has Current Version	<a href="http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/1/">http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/1/</a>
PAV Version ()	1
PAV Authored On ()	2009-11-03 16:19:38.0
Deprecated()	false
Broader	<a href="http://vocab.nerc.ac.uk/collection/P02/current/SIGT/">http://vocab.nerc.ac.uk/collection/P02/current/SIGT/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/S26/current/MAT00640/">http://vocab.nerc.ac.uk/collection/S26/current/MAT00640/</a>

e.g., <http://vocab.nerc.ac.uk/collection/P02/current/SIGT/>

↑ -- **Density of the water column --**

URI	<a href="http://vocab.nerc.ac.uk/collection/P02/current/SIGT/">http://vocab.nerc.ac.uk/collection/P02/current/SIGT/</a>
Identifier ()	SDN:P02::SIGT
Preferred label (en)	<b>Density of the water column</b>
Alternative label ()	WC_Dens
Version Info ()	1
Has Current Version	<a href="http://vocab.nerc.ac.uk/collection/P02/current/SIGT/1/">http://vocab.nerc.ac.uk/collection/P02/current/SIGT/1/</a>
PAV Version ()	1
PAV Authored On ()	2004-08-16 11:31:09.0
Deprecated()	Absolute determinations of water column density plus parameter
Broader	<a href="http://vocab.nerc.ac.uk/collection/P03/current/D020/">http://vocab.nerc.ac.uk/collection/P03/current/D020/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/D01/current/D0100001/">http://vocab.nerc.ac.uk/collection/D01/current/D0100001/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/L19/current/005/">http://vocab.nerc.ac.uk/collection/L19/current/005/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/L04/current/L04001/">http://vocab.nerc.ac.uk/collection/L04/current/L04001/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/P05/current/014/">http://vocab.nerc.ac.uk/collection/P05/current/014/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/P22/current/28/">http://vocab.nerc.ac.uk/collection/P22/current/28/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/">http://vocab.nerc.ac.uk/collection/P01/current/SIGTSG01/</a>
Broader	<a href="http://vocab.nerc.ac.uk/collection/P07/current/BBAH2104/">http://vocab.nerc.ac.uk/collection/P07/current/BBAH2104/</a>



# Controlled vocabularies in our domain

e.g., Climate and forecast (CF) metadata convention

<http://cfconventions.org/standard-names.html>

▶ <a href="#">sea_surface_wind_wave_period_at_variance_spectral_density_maximum</a>	s
▼ <a href="#">sea_water_density</a> Sea water density is the in-situ density (not the potential density). If 1000 kg m <sup>-3</sup> is subtracted, the standard name sea_water_sigma_t should be chosen instead.	kg m <sup>-3</sup>
▶ <a href="#">sea_water_neutral_density</a>	kg m <sup>-3</sup>

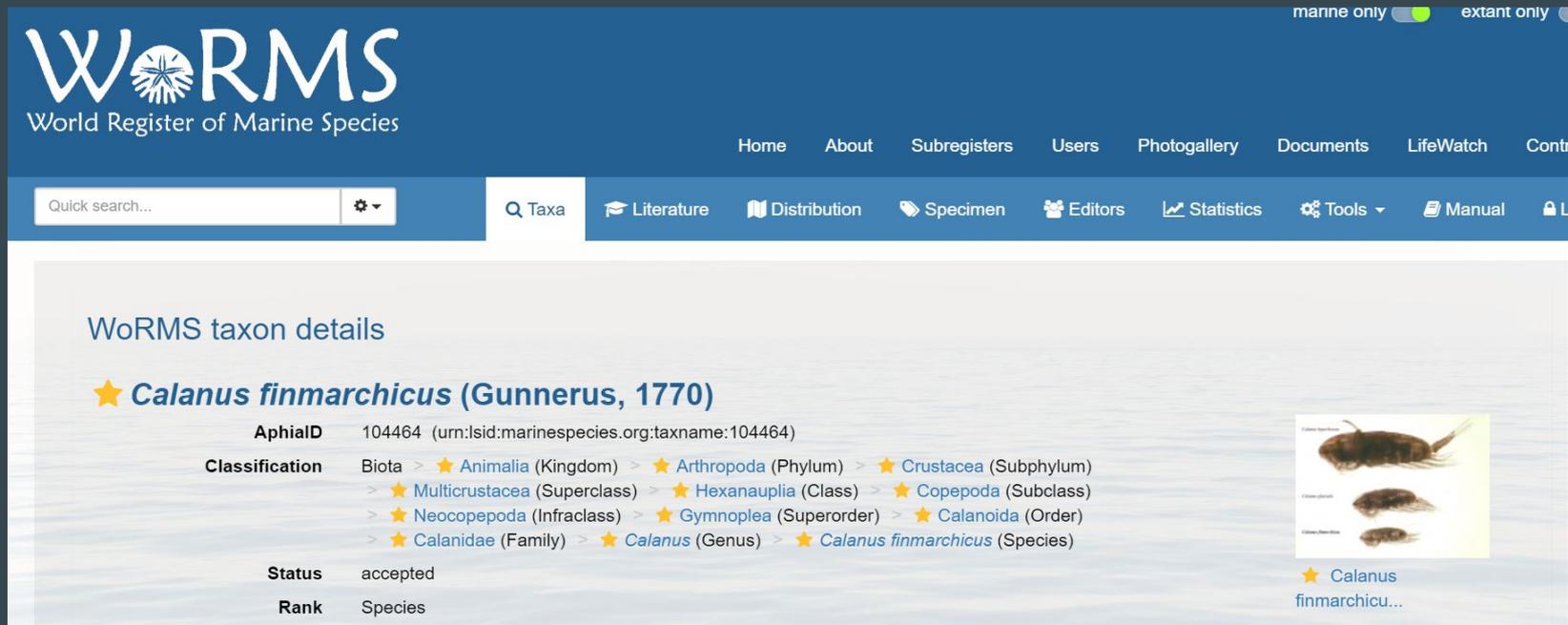
javascript:void(0)

Other examples:

- `acoustic_signal_roundtrip_travel_time_in_sea_water`
- `mass_concentration_of_nitrous_oxide_in_air`

# Taxonomy - WoRMs

The World Register of Marine Species (WoRMs) provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy.



The screenshot shows the WoRMS website interface. At the top, there is a blue header with the WoRMS logo and the text "World Register of Marine Species". To the right of the logo, there are two toggle switches: "marine only" (which is turned on) and "extant only". Below the header is a navigation menu with links for Home, About, Subregisters, Users, Photogallery, Documents, LifeWatch, and Contact. A search bar is located on the left side of the navigation menu. Below the navigation menu is a secondary menu with icons and labels for Taxa, Literature, Distribution, Specimen, Editors, Statistics, Tools, Manual, and a lock icon. The main content area is titled "WoRMS taxon details" and features a star icon next to the species name *Calanus finmarchicus* (Gunnerus, 1770). Below the species name, there are several fields: "AphiaID" with the value 104464 (urn:Isid:marinespecies.org:taxname:104464), "Classification" with a hierarchical list of taxonomic ranks from Biota to Species, "Status" with the value "accepted", and "Rank" with the value "Species". To the right of the classification and status fields, there is a small image showing three specimens of *Calanus finmarchicus* at different stages of development. Below the image, there is a star icon and the text "Calanus finmarchicu...".

WoRMS taxon details

★ *Calanus finmarchicus* (Gunnerus, 1770)

**AphiaID** 104464 (urn:Isid:marinespecies.org:taxname:104464)

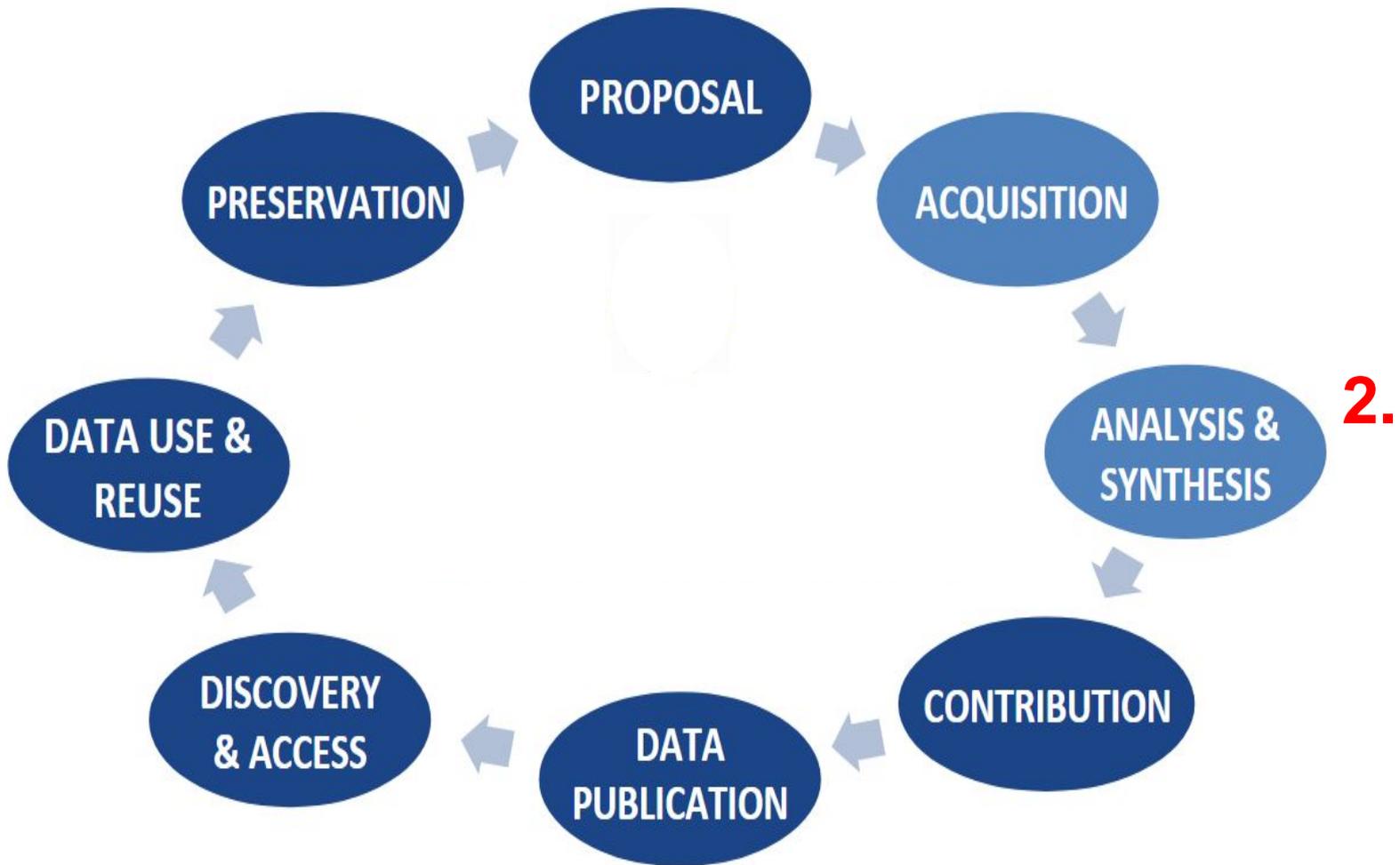
**Classification** Biota > ★ Animalia (Kingdom) > ★ Arthropoda (Phylum) > ★ Crustacea (Subphylum) > ★ Multicrustacea (Superclass) > ★ Hexanauplia (Class) > ★ Copepoda (Subclass) > ★ Neocopepoda (Infraclass) > ★ Gymnoplea (Superorder) > ★ Calanoidea (Order) > ★ Calanidae (Family) > ★ Calanus (Genus) > ★ *Calanus finmarchicus* (Species)

**Status** accepted

**Rank** Species

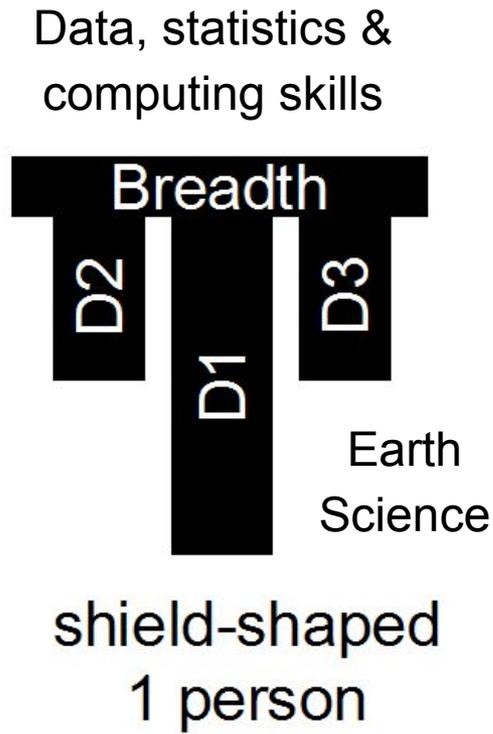
★ Calanus finmarchicu...

We recommend checking species names in WoRMs and including identifiers in your data when possible.

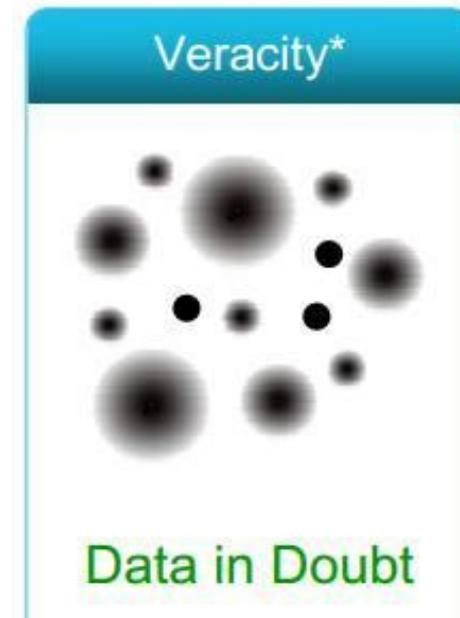


## THE DATA LIFE CYCLE

# We store, analyze, and present data... what else makes us different than data scientists outside our domain?



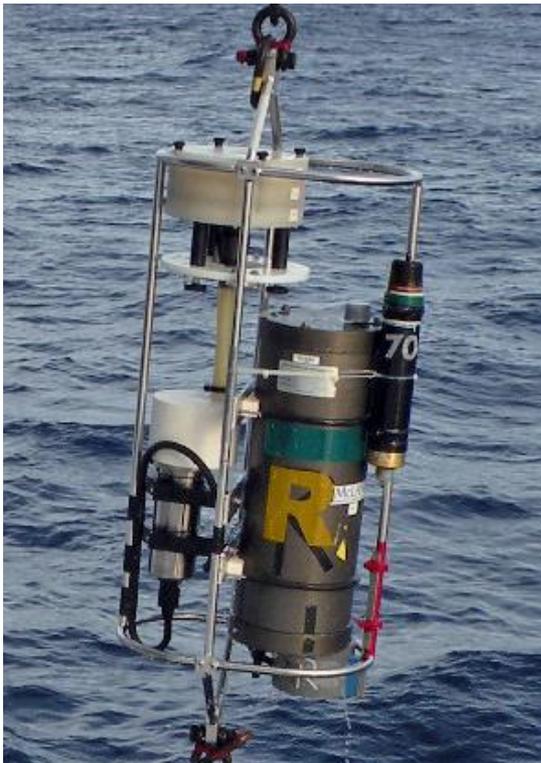
Deep knowledge of  
how to appropriately  
use the data



# We have knowledge and/or intuition about data quality and data uncertainty

To distinguish these concepts, watch Carol Anne Clayson's plenary talk "Information about the data is as important as the data itself" from 2017 ESIP Summer Meeting, 3:22 to 3:34 in:

<https://www.youtube.com/watch?v=8fP4M0iAYGs>



We get to know our instruments,  
e.g., "Ringo"



# Quality Flags

You can incorporate data quality and uncertainty into your data and/or metadata.

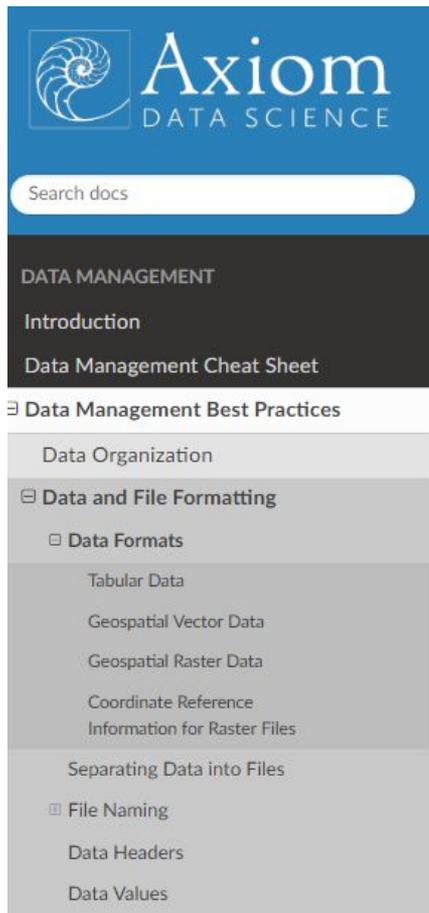
e.g., International Oceanographic Data and Information Exchange (IODE) quality flags.

See: [https://www.iode.org/mg54\\_3](https://www.iode.org/mg54_3)

Value	Primary-level flag short name	Definition
1	Good	Passed documented required QC tests
2	Not evaluated, not available or unknown	Used for data when no QC test performed or the information on quality is not available
3	Questionable/suspect	Failed non-critical documented metric or subjective test(s)
4	Bad	Failed critical documented QC test(s) or as assigned by the data provider
9	Missing data	Used as place holder when data are missing

# We encounter a variety of data formats in our domain

<https://www.axiomdatascience.com/best-practices/DataandFileFormatting.html#data-formats>



The screenshot shows the Axiom Data Science website navigation menu. At the top is the Axiom Data Science logo, which includes a blue square with a white nautilus shell icon and the text 'Axiom DATA SCIENCE'. Below the logo is a search bar with the placeholder text 'Search docs'. The main navigation menu is a dark blue bar with white text, containing 'DATA MANAGEMENT' and 'Introduction'. Below this is a 'Data Management Cheat Sheet' link. A sub-menu is open, showing 'Data Management Best Practices' with a dropdown arrow. The sub-menu items are: 'Data Organization', 'Data and File Formatting' (with a dropdown arrow), 'Data Formats' (with a dropdown arrow), 'Tabular Data', 'Geospatial Vector Data', 'Geospatial Raster Data', 'Coordinate Reference Information for Raster Files', 'Separating Data into Files', 'File Naming' (with a dropdown arrow), 'Data Headers', and 'Data Values'.

## Data and File Formatting

### Data Formats

In choosing a file format, data collectors should select a format that is useable, open, and that will likely be readable well into the future. Microsoft Excel, as an example, is a useful tool for data manipulations and data visualization, but versions of Excel files may become obsolete and may not be easily readable over the longer term. Likewise, database management systems (DBMS) like MS Access, Filemaker Pro, and others, can be a very effective way to store and query data, but the raw formats tend to change over time (even a few years). If your program or organization has used these or other proprietary DBMS tools, it is essential to plan for exporting your data in a stable, well-documented, and non-proprietary format.

Below is a summary of the suggested tabular, image, and GIS data file formats suitable for long-term archiving.

- Containers: TAR, GZIP, ZIP
- Databases: CSV, XML
- Tabular data: **CSV**
- Geospatial vector data: SHP, GeoJSON, KML, DBF, NetCDF
- Geospatial raster data: GeoTIFF/TIFF, **NetCDF**, HDF-EOS
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

**We are highlighting CSV and NetCDF in this workshop**

# Data formats in our domain

e.g., NetCDF (<https://www.unidata.ucar.edu/software/netcdf/>)



The screenshot shows the Unidata website interface. At the top, there is a navigation menu with links for Data, Software, Downloads, Support, Community, Projects, News, Events, and About Us. Below the menu is the Unidata logo, which includes the UCAR Community Programs logo and the text "unidata Data Services and Tools for Geoscience". To the right of the logo is a search bar labeled "Google Custom Search" with a magnifying glass icon. Below the search bar is a breadcrumb trail: "Unidata Home » NetCDF". The main content area is titled "Network Common Data Form (NetCDF)". On the left side, there is a sidebar with links for "NETCDF", "Release Notes", "FAQs", and "Documentation". The main content area contains a description of NetCDF: "NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data." Below this description is a small icon of the netCDF logo and a link that says "See the netCDF package overview ▶".

NetCDF (.nc) is a standard delivery format for OOI, PO.DAAC, BODC, and other repos in our domain

NCEI templates: <https://www.nodc.noaa.gov/data/formats/netcdf/v2.0/>

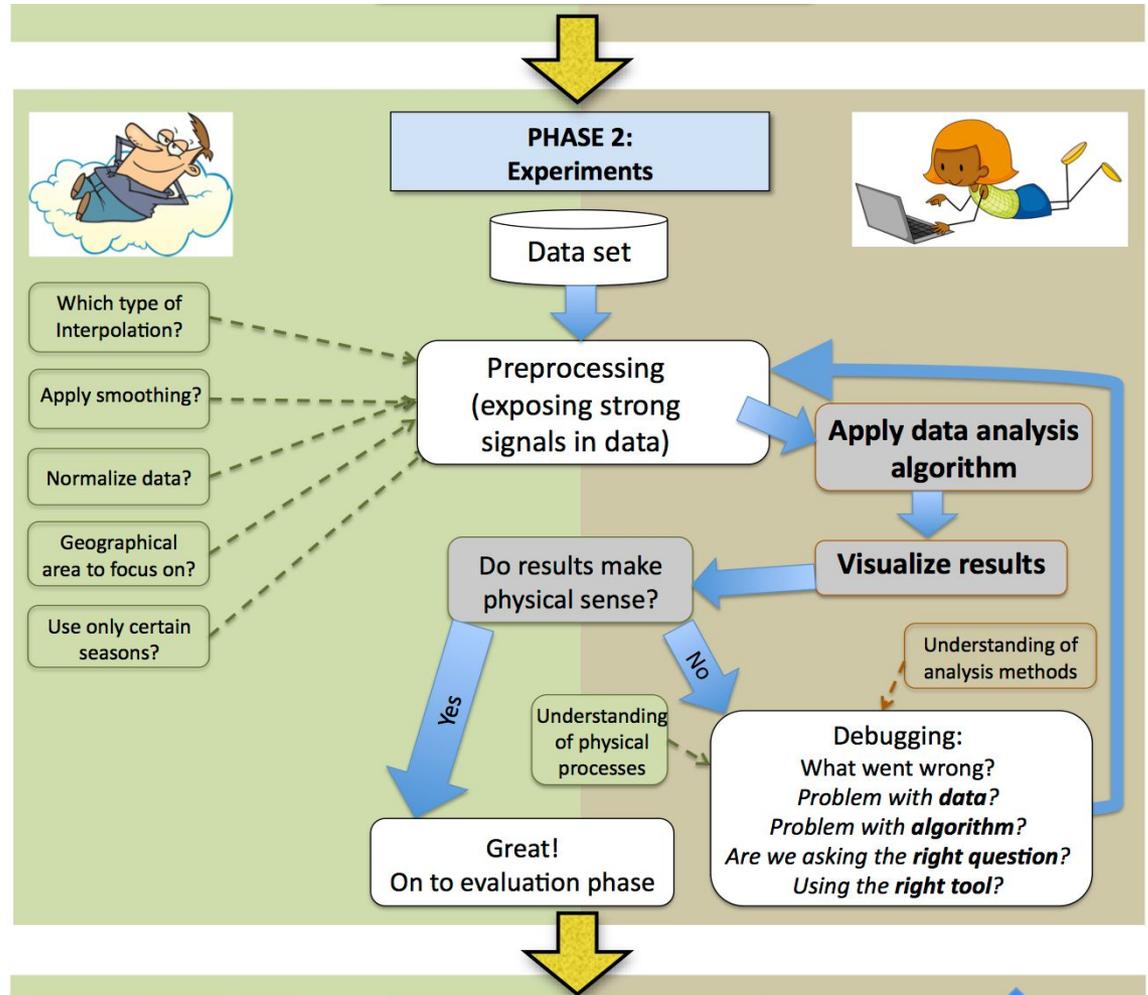
The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF API.

<http://cfconventions.org/Data/cf-documents/overview/viewgraphs.pdf>

# Three Steps to Successful Collaboration with Data Scientists

“The ability to store and exchange data in a standardized way is an essential requirement to build the data processing e-infrastructure needed for Big Data analysis.”

[EU Marine Board Future Science Brief](#)



# Types of Analyses

- Processing: Subsetting, Merging, Manipulating
- Statistical analyses
- Graphical analyses and visualization



## Data Analysis and Workflows:

[https://dataoneorg.github.io/Education/lessons/09\\_analysis/slides.html#1](https://dataoneorg.github.io/Education/lessons/09_analysis/slides.html#1)

Earth Science Data Analytics: "The process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data, encompassing varieties of data types... to uncover patterns, correlations, and other information, to better understand our Earth."

**When analyzing our own data, we often want to compare to data in other studies.**

**Were you able to find/access the data reported in your homework paper?**

# We encounter a variety of types of software to analyze data in our domain

e.g.,



## ArcGIS

Fledermaus



Ocean Data View

[Home](#) [Data](#) [Software](#) [Documentation](#) [Links](#) [ODV Forum](#)

geneious  
prime

# We use and create software to process/analyze/visualize data in our domain

How do you find out about new software that you might want to use?

Results of poll at the 2018 AGU Fall Meeting Town Hall for sharing scientific software:

people *either* found software in a Google search *or* they found out within their domain \* (asked their colleagues, learned in papers, saw at conferences, etc.)

\* this is why we have tables-by-department today for our discussion

# Were you able to find/access the software used in your homework paper?

***Take the next 10 min*** at your table to discuss with your peers the types of analyses in your homework paper, and whether you could access the software for those analyses

Add the journal for your homework paper to our shared notes

“welcoming and supportive environment for all”

# Follow-up to homework

<http://scientificpaperofthefuture.org/gpf/>

## Scientific Paper of the Future

### Modern Paper

#### Text:

Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned

#### Data:

Stall, S., et al. (2019) Make scientific data FAIR. *Nature*, 570 (7759): 27, [doi:10.1038/d41586-019-01720-7](https://doi.org/10.1038/d41586-019-01720-7)

### Reproducible Publication

#### Software:

For data preparation, data analysis, and visualization

#### Provenance and methods:

Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies

### Open Science

#### Sharing:

Deposit data and software (and provenance/workflow) in publicly shared repositories

#### Open licenses:

Open source licenses for data and software (and provenance/workflow)

#### Metadata:

Structured descriptions of the characteristics of data and software (and provenance/workflow)

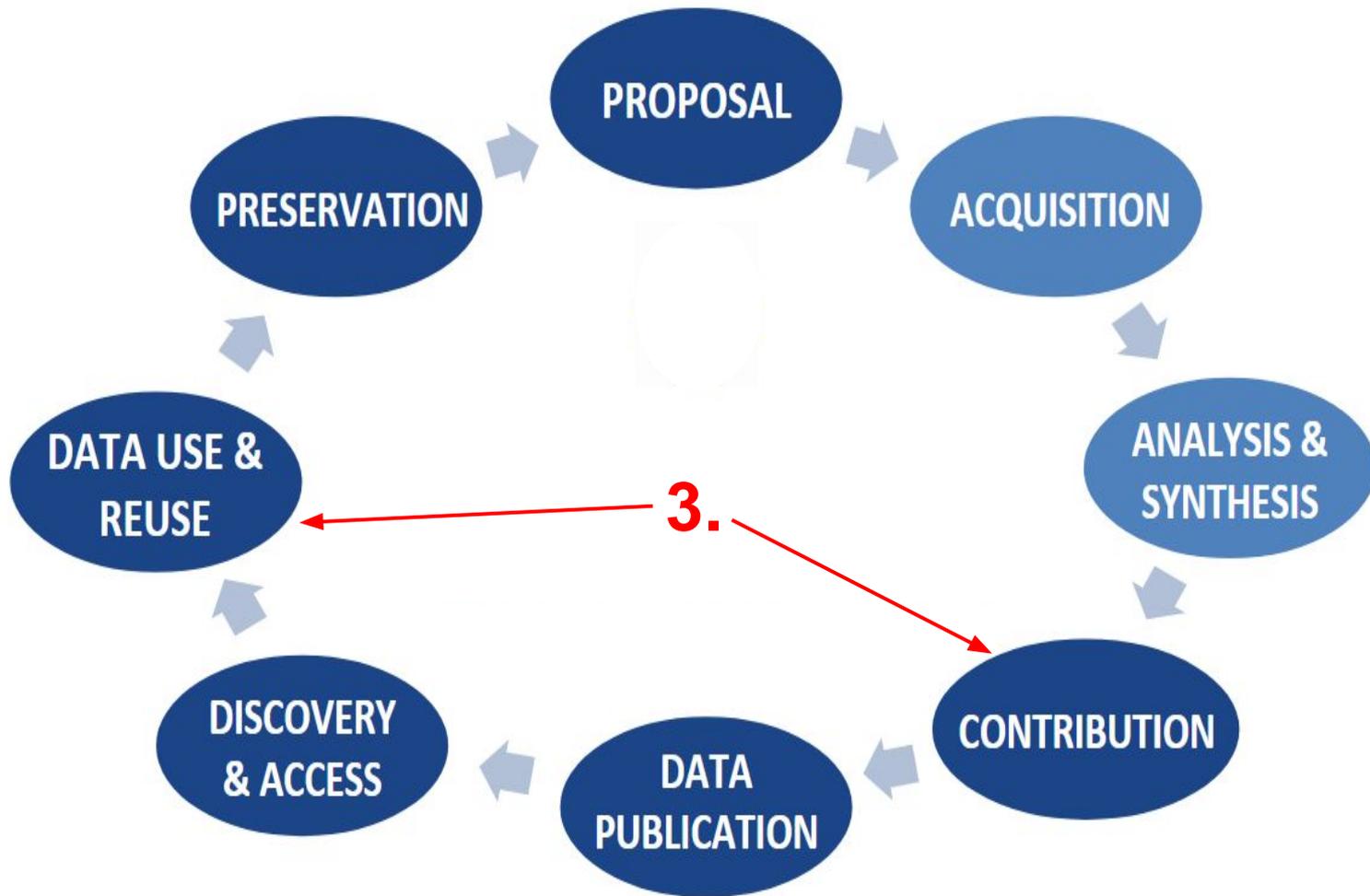
### Digital Scholarship

#### Persistent identifiers:

For data, software, and authors (and provenance/workflow)

#### Citations:

Citations for data and software (and provenance/workflow)



## THE DATA LIFE CYCLE

**“A scholar’s positive contribution is measured by the sum of the original data that [s/he] contributes. Hypotheses come and go but data remain.”**

from: Advice to a Young Investigator (Santiago Ramón y Cajal, 1897)



*photo by Chris Linder (WHOI)*

# Results from yesterday's poll: Repositories that you have contributed to

NCEI:

National Geophysical Data Center (NGDC) and  
Ocean Carbon Data System (OCADS)

Arctic Data Center (was ACADIS)

EDI

IRMA (National Parks Service)

Zenodo

# How easy or difficult was the process of contributing data to a repository?

- Very easy
- Relatively easy
- Neutral
- Relatively difficult
- Very difficult

# Why share your data?

## Benefit to...

### YOU:

- Credit for your hard work!
- Can increase citations
- Can boost collaborations
- Increases exposure
- Satisfies funder requirements

### Research Community:

- Builds a community resource
- Enables new discoveries
- Sparks collaborations
- Allows for transparency and reproducibility of research results



### Society at-large:

- Transparency boosts public confidence in scientific process
- Can contribute to management and policy
- Availability to audiences outside of research (education, general public)



# 7 HABITS FOR SHARING YOUR DATA WITH REPOSITORIES

1. **Start Early!**
2. Communicate with Collaborators
3. Connect with Funders
4. Seek out Available Resources
5. Reach out to Repo
6. Document in a DMP
7. Follow Through! (with data best practices)

# FUNDER -- REPO -- RESEARCHER



National Science Foundation  
WHERE DISCOVERIES BEGIN



Geosciences (GEO)

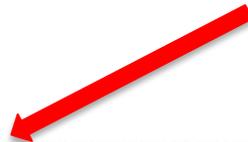


Email Print Share

## OCE Policies on Data Management, Ship Time, Resubmission and Environmental Compliance

June 26, 2017

1. **NSF Proposal & Award Policies & Procedures Guide (PAPPG) including Guidance on Letters of Collaboration for Unfunded Collaborations**
2. **Data Management**
  - [NSF Dissemination and Sharing of Research Results](#)
  - [Division of Ocean Sciences \(OCE\) Sample and Data Policy](#) including preferred data and physical collection archives and centers
  - [Biological and Chemical Oceanography Data Management Office \(BCO-DMO\)](#): BCO-DMO has developed a [Data Management Plan template](#) to assist investigators in submission of plans that meet the NSF OCE Sample and Data Policy requirements.
  - [Interdisciplinary Earth Data Alliance \(IEDA\)](#): IEDA [Data Management Plan \(DMP\) Tool](#) provides an easy way to generate a data management plan for inclusion in NSF proposals.
3. **Proposals That Include Ship Time**
  - [OCE Guidance for Proposals that Include Ship Time](#)
  - [UNOLS Cruise Planning Information for Scientists](#)



# LOST?

- What to do in the Absence of Guidance?



# HABITS...

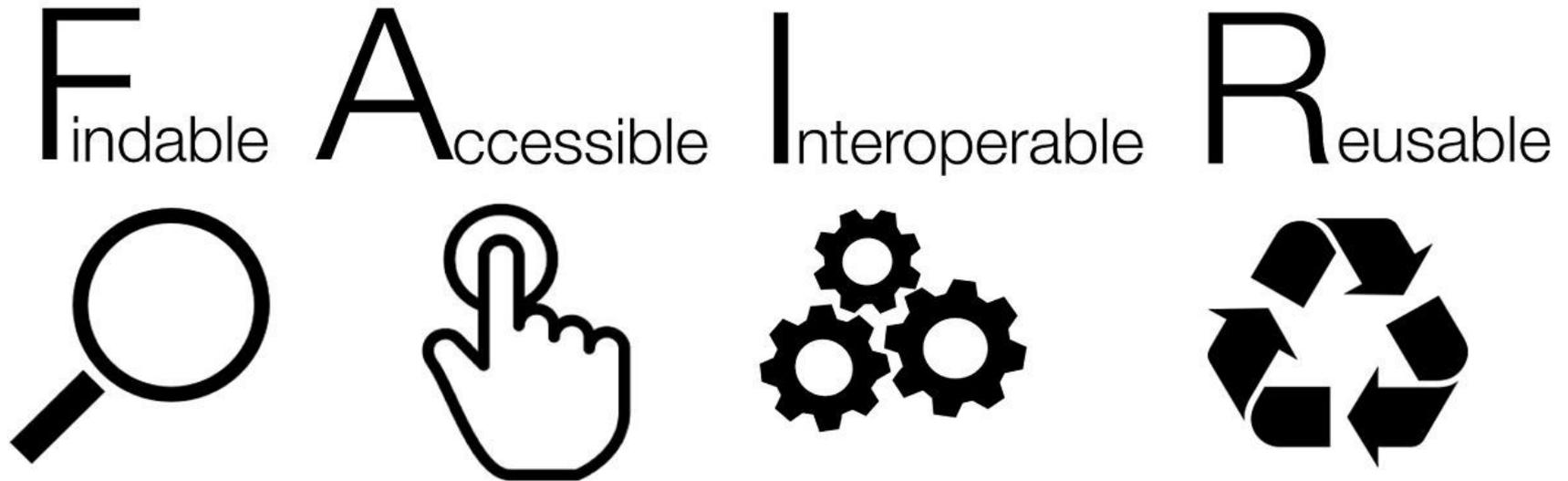
1. Domain
2. Institutional
3. Generic



Module 2: Finding Repositories (re3data)

<https://www.re3data.org/>

# Recall the goal to make your data FAIR...

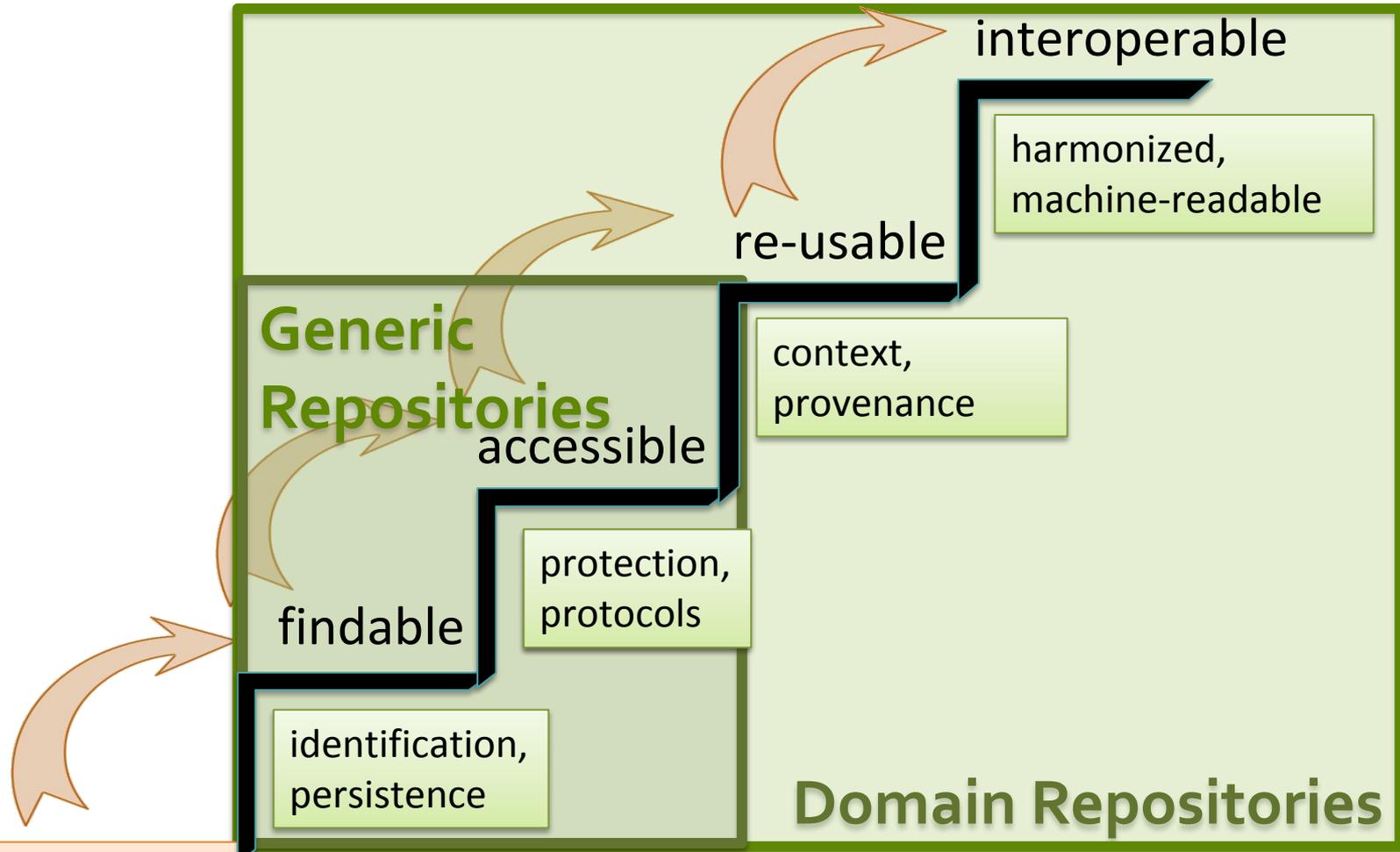


Original article describing FAIR data principles: Wilkinson, M.D., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018, <https://doi.org/10.1038/sdata.2016.18>

Article more specific to FAIR in our domain: Stall, S., et al. (2018), Advancing FAIR data in Earth, space, and environmental science, *Eos*, 99, <https://doi.org/10.1029/2018EO109301>

# HABITS...

FAIR Data



PI Managed Data

# Activity for the “A”, “I”, and “R” in FAIR

## What is ERDDAP? (simple)

- ERDDAP is a NOAA developed **data server**
- middle-man that facilitates **open** and **accessible** data download in a variety of formats applicable to various communities.
- **Accepts a variety of file formats** on the backend (ASCII text, netCDF, compressed files, image files, audio files, and others)
- Allows for **data and metadata** coupling and download
- Provides **web accessible urls** for advanced data download and subsetting via other software packages

Slide from “BCO-DMO’s migration to ERDDAP,” DOI:10.1575/1912/24933, <https://hdl.handle.net/1912/24933>

“ERDDAP used to be an acronym, but it outgrew that original description. Now, please just think of it as a name, not an acronym.” <https://coastwatch.pfeg.noaa.gov/erddap/information.html>

# Let's pull some data from BCO-DMO's ERDDAP directly into our software environment

Search BCO-DMO's ERDDAP for Dennis McGillicuddy (AOPE Chair)

<https://erddap.bco-dmo.org/erddap/index.html>

e.g.,

[https://erddap.bco-dmo.org/erddap/taledap/bcodmo\\_dataset\\_3358.html](https://erddap.bco-dmo.org/erddap/taledap/bcodmo_dataset_3358.html)



**BCO-DMO ERDDAP**  
Accessing BCO-DMO data

## ERDDAP > search

Do a Full Text Search for Datasets:

2 matching datasets, with the most relevant ones listed first.  
(Or, refine this search with [Advanced Search](#))

Grid DAP Data	Sub-set	Table DAP Data	Make A Graph	W M S	Source Data Files	Access-ible	Title	Sum-mary	ISO, Metadata	Back-ground Info	RSS	Institution	Dataset ID
		data	graph		files	public	Pigments from HPLC analysis of bottle samples collected aboard Oceanus from R/V Oceanus OC404-01, OC404-04, OC415-01, OC415-03 in the Sargasso Sea from 2004-2005 (EDDIES project)		I M	<a href="#">background</a>		BCO-DMO	bcodmo_dataset_3024
		data	graph		files	public	Bottle Data from multiple cruises in the Gulf of Maine, NA4, 43 30N, 69 00W, Gulf of Maine, Mass Bay to Bay of Fundy, Cape Cod Bay, 2003-2010 (ALEX-GoME project)		I M	<a href="#">background</a>		BCO-DMO	bcodmo_dataset_3358

# Let's pull some data from BCO-DMO's ERDDAP

Click on "data"  
to generate URL

or graph  
to explore in ERDDAP

ERDDAP > tabledap > Data Access Form

Dataset Title: **Bottle Data from multiple cruises in the Gulf of Maine, NA4, 43 30N, 69 00W, Gulf of Maine, Mass Bay to Bay of Fundy, Cape Cod Bay, 2003-2010 (ALEX-GoME project)**

Institution: BCO-DMO (Dataset ID: bcodmo\_dataset\_3358)  
Information: Summary | License | ISO 19115 | Metadata | Background | Files | Make a graph

Variable	Check All	Uncheck All	Optional Constraint #1	Optional Constraint #2	Minimum	Maximum
<input checked="" type="checkbox"/> Year (unitless)			>=	<=	2003	2010
<input checked="" type="checkbox"/> cruise_id (text)			>=	<=		
<input checked="" type="checkbox"/> station (x.x)			>=	<=	1.0	257.0
<input checked="" type="checkbox"/> date (unitless)			>=	<=		
<input checked="" type="checkbox"/> time2 (Time, HHMM)			>=	<=		
<input checked="" type="checkbox"/> latitude (degrees_north)			>=	<=	40.052	45.095833
<input checked="" type="checkbox"/> longitude (degrees_east)			>=	<=	-71.194833	-65.728
<input checked="" type="checkbox"/> depth (m)			>=	<=	1.0	570.0
<input checked="" type="checkbox"/> Alex (cells/L)			>=	<=	-9.99	38565.0
<input checked="" type="checkbox"/> Alex_Live (cells/L)			>=	<=	-9.99	11662.0
<input checked="" type="checkbox"/> Pressure (decibars)			>=	<=	-9.99	570.288
<input checked="" type="checkbox"/> Temperature (degrees Celcius)			>=	<=	-9.99	21.716
<input checked="" type="checkbox"/> Salinity (psu)			>=	<=	-9.99	35.809
<input checked="" type="checkbox"/> NO3_plus_NO2 (uM)			>=	<=	-9.99	24.14
<input checked="" type="checkbox"/> Silicate (uM)			>=	<=	-9.99	33.28
<input checked="" type="checkbox"/> NH4 (uM)			>=	<=	-9.99	16.773
<input checked="" type="checkbox"/> PO4 (uM)			>=	<=	-9.99	4.72
<input checked="" type="checkbox"/> Chla (ug/l)			>=	<=	-9.99	22.79
<input checked="" type="checkbox"/> Phaeo (ug/l)			>=	<=	-9.99	8.68
<input checked="" type="checkbox"/> Flag (Integer)			>=	<=	1	5

Server-side Functions

distinct()

File type: (more info)

.htmlTable - View a UTF-8 .html web page with the data in a table. Times are ISO 8601 strings.

Just generate the URL:

(Documentation / Bypass this form)

Submit (Please be patient. It may take a while to get the data.)

Choose csvp (or .nc)

BCO-DMO BCO-DMO ERDDAP  
Biological & Chemical Oceanography Data Management Office  
Accessing BCO-DMO data

ERDDAP > tabledap > Make A Graph

Dataset Title: **Bottle Data from multiple cruises in the Gulf of Maine, NA4, 43 30N, 69 00W, Gulf of Maine, Mass Bay to Bay of Fundy, Cape Cod Bay, 2003-2010 (ALEX-GoME project)**

Institution: BCO-DMO (Dataset ID: bcodmo\_dataset\_3358)  
Range: longitude = -71.19483 to -65.728°E, latitude = 40.052 to 45.095833°N, depth = 1.0 to 570.0m  
Information: Summary | License | ISO 19115 | Metadata | Background | Data Access Form | Files

Graph Type: markers

X Axis: longitude  
Y Axis: latitude  
Color: Year

Click on the map to specify a new center point.

Zoom: Out 8x | Out 2x | Out | Data | In | In 2x | In 8x

Constraints	Optional Constraint #1	Optional Constraint #2
>=	>=	<=
>	>	<
>=	>=	<=
>	>	<

Server-side Functions

distinct()

Graph Settings

Marker Type: Filled Square Size: 5  
Color:   
Color Bar: Continuity: Scale: N Sections:   
Minimum: Maximum:   
Draw land mask:   
Y Axis Minimum: Maximum: Ascending: ascending

Redraw the Graph (Please be patient. It may take a while to get the data.)

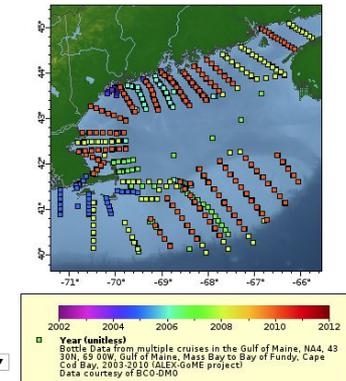
Optional:

Then set the File Type: .htmlTable (File Type information)

and Download the Data or an Image

or view the URL: [https://erddap.bco-dmo.org/erddap/tabledap/bcodmo\\_dataset\\_3358.htmlTab](https://erddap.bco-dmo.org/erddap/tabledap/bcodmo_dataset_3358.htmlTab)

(Documentation / Bypass this form)



[https://erddap.bco-dmo.org/erddap/tabledap/bcodmo\\_dataset\\_3358.html](https://erddap.bco-dmo.org/erddap/tabledap/bcodmo_dataset_3358.html)

# Take the next 10 min to access data via Application Programming Interface (API)

ERDDAP URL pattern

<https://namespace/erddap/tabledap/datasetID.fileType{?query}>

Example URL:

[https://erddap.bco-dmo.org/erddap/tabledap/bcodmo\\_dataset\\_3358.csvp?Year%2Ccruise\\_id%2Cstation%2Cdate%2Ctime%2Clatitude%2Clongitude%2Cdepth%2CAlex\\_Live%2CTemperature%2CNO3\\_plus\\_NO2%2CChla](https://erddap.bco-dmo.org/erddap/tabledap/bcodmo_dataset_3358.csvp?Year%2Ccruise_id%2Cstation%2Cdate%2Ctime%2Clatitude%2Clongitude%2Cdepth%2CAlex_Live%2CTemperature%2CNO3_plus_NO2%2CChla)



```
DennisData <- read.csv('URL')  
View(DennisData)
```



```
import pandas as pd  
DennisData = pd.read_csv('URL')  
DennisData.head(5)
```



```
options = weboptions('ContentType', 'table');  
DennisData = webread('URL', options);
```

[Click for Matlab NetCDF](#) (slide 6)

# What do the “I” and “R” in FAIR mean to you?

Did anyone pull data into an analysis tool?

Did you do this by manual download or API?

Did anyone inspect the metadata for interoperability?

Do you think you could reuse the data?

Check out data access options in other domain repositories,

e.g.,:

[PO.DAAC](#)

[EDI](#)

Follow-up article for ERDDAP activity:

Signell, R.P.; Fernandes, F.; Wilcox, K. (2016) Dynamic Reusable Workflows for Ocean Science. *J. Mar. Sci. Eng.*, 4, 68, <https://doi.org/10.3390/jmse4040068>

# What about the “F” Findable and ERDDAP?

ERDDAP megasearch:

<http://erddap.com/>

ERDDAP Dataset Discovery 

Search Datasets

McGillicuddy

Type some words about the dataset you seek, then press the green button

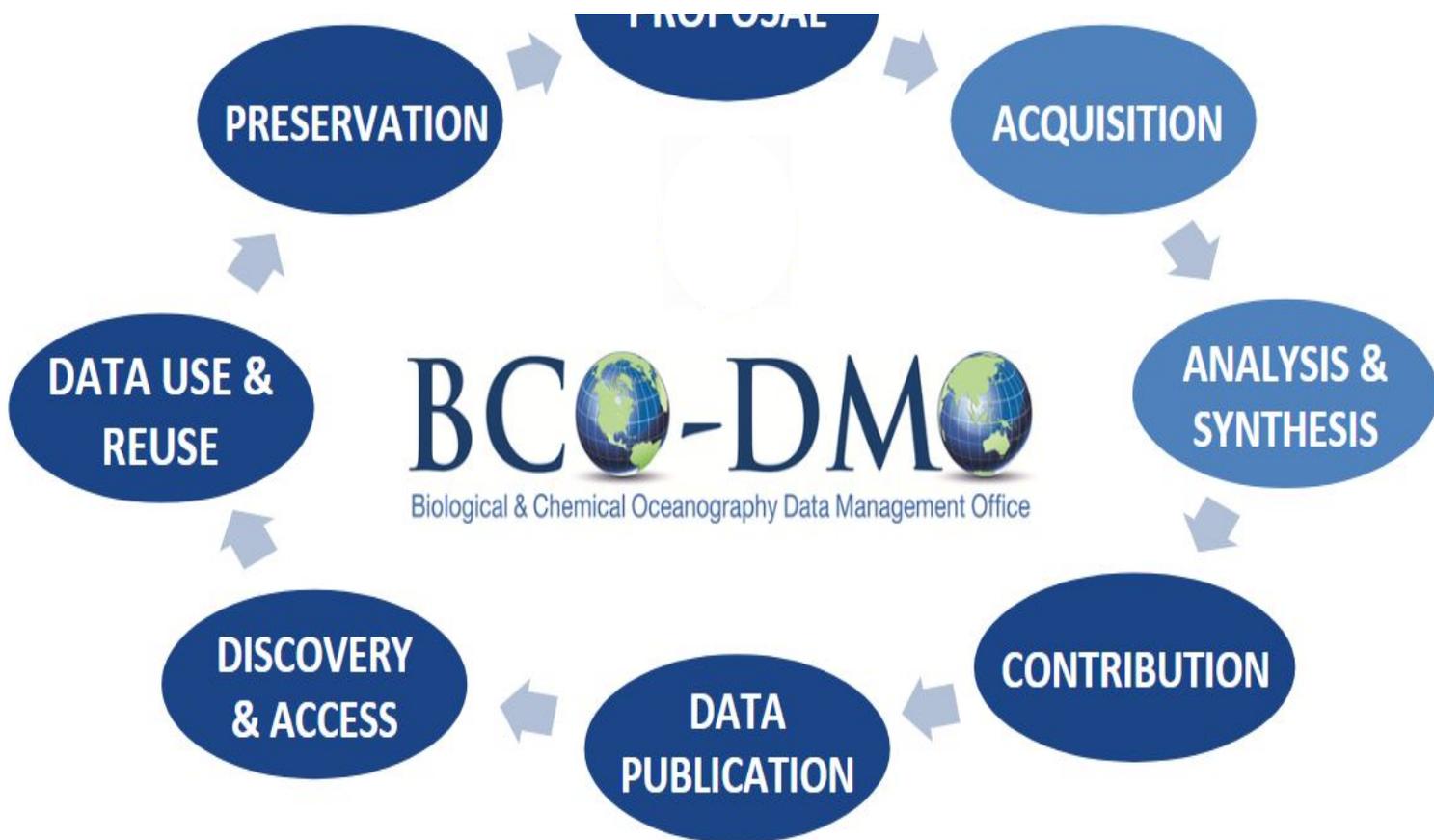
Searched 44 ERDDAP servers; found 3 datasets from 1 server; total search time 5971ms.

Title	Institution	Dataset
 Pigments from HPLC analysis of bottle samples collected aboard Oceanus from R/V Oceanus OC404-01, OC404-04, OC415-01, OC415-03 in the Sargasso Sea from 2004-2005 (EDDIES project)	BCO-DMO	<a href="http://erddap.bco-dmo.org/bcodmo_dataset_3024">bcodmo_dataset_3024</a> erddap.bco-dmo.org
 Bottle Data from multiple cruises in the Gulf of Maine, NA4, 43 30N, 69 00W, Gulf of Maine, Mass Bay to Bay of Fundy, Cape Cod Bay, 2003-2010 (ALEX-GoME project)	BCO-DMO	<a href="http://erddap.bco-dmo.org/bcodmo_dataset_3358">bcodmo_dataset_3358</a> erddap.bco-dmo.org
 Mesozooplankton biomass from MOCNESS tows collected during R/V Oceanus cruises OC415-01, OC415-03, OC404-01, OC404-04 in the Sargasso Sea, 2004-2005 (EDDIES project)	BCO-DMO	<a href="http://erddap.bco-dmo.org/bcodmo_dataset_3211">bcodmo_dataset_3211</a> erddap.bco-dmo.org

<https://coastwatch.pfeg.noaa.gov/erddap/download/SearchMultipleERDDAPs.html>

<https://github.com/IrishMarineInstitute/awesome-erddap>

# Congratulations! You completed WHOI's Data Science Training Camp! What's next?



THE DATA LIFE CYCLE

# Strategies to learn more

- Self-paced learning (online resources, books)
- Join and/or create communities
- Attend workshops in person

See Box 1 and Box 2 in:

Lowndes, J.S.S., et al. (2017) Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, 1, 0160, doi:10.1038/s41559-017-0160

# Online resources: Data life cycle

- ESIP Data Management Training (DMT) Clearinghouse  
<http://dmtclearinghouse.esipfed.org/>
- DataONE Education Modules  
<https://www.dataone.org/education>
- Data Carpentry lessons  
<https://datacarpentry.org/lessons/>

**Plus resources listed in  
Module 2 slides**

# Online resources: Software and computing

- Udemy

<https://www.udemy.com/>

- Coursera

<https://www.coursera.org/browse/data-science>

- Software Carpentry lessons

<https://software-carpentry.org/lessons/>

# Online resources tailored to our domain

- International Oceanographic Data and Information Exchange (IODE) OceanTeacher Global Academy  
<https://classroom.oceanteacher.org/> (materials based on in-person workshops that you can apply to)
- OceanHackWeek:  
<https://github.com/oceanhackweek/Oceans19-data-science-tutorial>
- Integrated Marine Observing System (IMOS) (Australia)  
Marine Data and Science e-lectures  
<https://open2u.utas.edu.au/Course/4261>

# Resources at MIT and in Boston area

- MIT Library is offering workshops, e.g., “Make your research computationally reproducible,” [Introduction to Python on Jan. 28th](https://libraries.mit.edu/news/events/) (<https://libraries.mit.edu/news/events/>)
- MIT Statistics and Data Science Center <https://stat.mit.edu/>  
SDSCon 2020, MIT Statistics and Data Science Conference, Friday, April 3, 2020 @E14-674.
- Introduction to Computational Thinking and Data Science  
<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/>

<https://www.meetup.com/Boston-useR/>



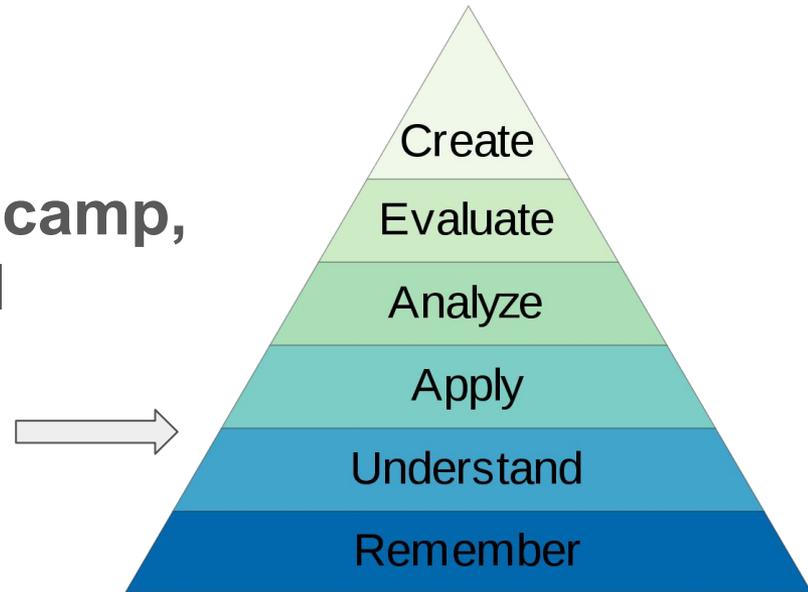
# Resources in our Woods Hole community

- Woods Hole data-mongers email list  
<http://mailman.who.edu/mailman/listinfo/data-mongers>
- WHOI Software email list  
<http://mailman.who.edu/mailman/listinfo/whoi-software>
- Bioinformatics email list  
<http://mailman.who.edu/mailman/listinfo/bioinformatics-users>
- WHOI's Matlab users group (mug@who.edu)
- Woods Hole Python users group (whython@googlegroups.com)
- There is interest but no champion yet:
  - local R users group
  - local data science discussion group (*start with your lab group*)

We are planning a Summer Series with 4 “Lunch-and-Learns”...

# “Exposure and Confidence” \*

**When you complete the camp,  
you may reach this level  
for some skills**



“Bloom's taxonomy” of levels of learning

**For higher level learning, we will offer  
The Carpentries Workshops...**

\* Lowndes, J.S.S., et al. (2017) Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, 1, 0160, doi:10.1038/s41559-017-0160

**Please choose name tag by department.  
Goal is for each table to have 1 or 2 depts.**



**AOPE**



**Biology**



**G&G**



**MC&G**



**PO**



**Other, or would rather  
not specify**