Data science training for today's oceanographers: Curriculum development across disciplines at Woods Hole Oceanographic Institution

Stace Beaulieu (stace@whoi.edu), Joe Futrelle, Andrei Huang, Danie Kinkade, Roberta Mazzoli, Audrey Mickle, Shannon Rauch, Lisa Raymond, & Nick Symmonds

Why develop a Data Science Training Camp?

It is imperative for today's oceanographers to have practical training in data science to support research transparency, reproducibility, and the sharing of data and software. In the past few years Woods Hole Oceanographic Institution (WHOI) has supported an increasing number of informal learning opportunities for data science skills through workshops and tutorials. In 2017 WHOI became a member in The Carpentries, an international organization with foundational data science training well-recognized across scientific disciplines. However, in addition to foundational skills, there is a need for learning best practices and skills in our ocean sciences domain.

Main goals for the camp:

- ✓ Foundational best practices in data, software, and project management for scientific research *(Wilson et al. 2017)*
- ✓ Some more specific practices in the ocean sciences
- ✓ Some resources available at WHOI
- ✓ Networking.

Target participants: We focused on WHOI/MIT Joint Program students and WHOI postdocs, but offered the training to other students and all technical and scientific staff in all departments. No prerequisites.

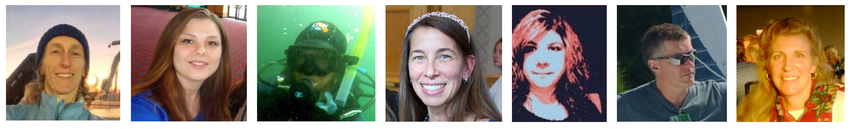
https://tinyurl.com/WHOI-Data-Science-syllabus

We delivered the curriculum in two half-day sessions each with two modules. The first session requires that participants from different departments sit together, while the second has participants from the same department sit together to discuss discipline-specific needs. Each module includes discussions and/or hands-on activities for inquiry-based learning in addition to prepared slides. Pre- and post-workshop surveys were conducted to assess participants' needs and learning attained. Live polls were conducted during each module, with a focus on the 1st session to provide input to the 2nd.

Instructors and helpers with wide range of technical expertise

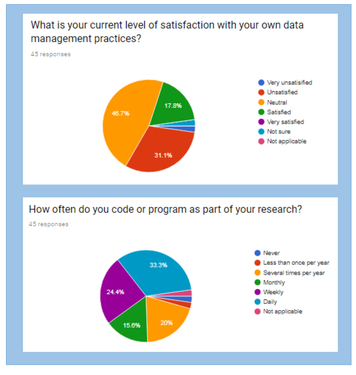
Scientists, Data Managers, Software Developers, Librarians, System Administrators

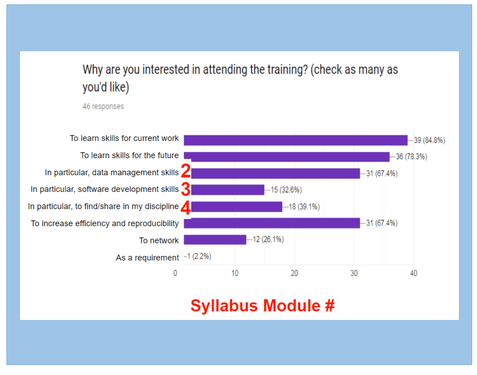
100+ years of experience shared in 8 hours!



Pre-Survey

We matched some of our pre-survey questions to those used by The Carpentries *(Jordan et al. 2018)*. We also collected some demographics to know that our participants were from all departments and all levels - from undergrads to senior scientists!

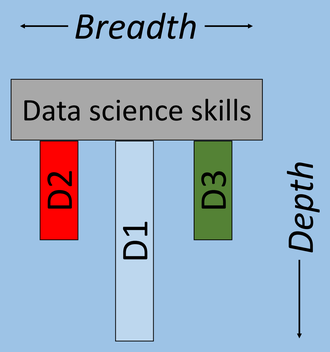




Shield-shaped model for the trainee

- ✓ Depth of knowledge in primary discipline
- ✓ Practical understanding of other disciplines
- ✓ Breadth of transferable skills

(Bosque-Pérez et al. 2016; Pennington et al. 2019)



Departments (Disciplines)

- Applied Ocean Physics and Engineering
- Biology
- Geology & Geophysics
- Marine Chemistry & Geochemistry
- Physical Oceanography
- Other, or would rather not specify

Implementation: 4 x 2-hour modules

Module 1

"Good enough practices in scientific computing"

Discuss challenges and solutions with participants from all disciplines:

- Data Management
- Software Development
- Project Management and Collaboration

Module 2

Data management

Introduce data life cycle and FAIR data principles

Hands-on activities examining tabular data and exploring WHOI's institutional repositories

Module 3

Software development and research computing

Demo several tools including Jupyter Notebooks and WHOIGit

Introduce WHOI IS resources including our high-performance computing (HPC) cluster

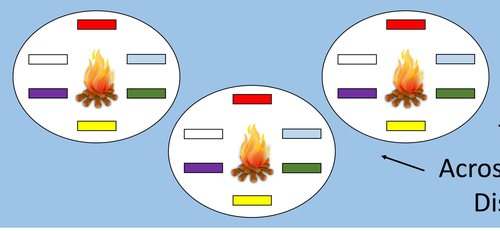
Live interactive notebooks: https://tinyurl.com/pandas-talk

Module 4

Best practices in the ocean sciences

Discuss homework with participants in same discipline

Use web-based API to access data directly into Matlab, R, and Python



Polls during the Camp

How comfortable are you with managing terabytes of data?

- *Relatively uncomfortable*

Have you submitted to a data repository?

- *No*

How familiar are you with version control software (e.g., git)?

- *Some knowledge*

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

- *Yes = ~ half the room*

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

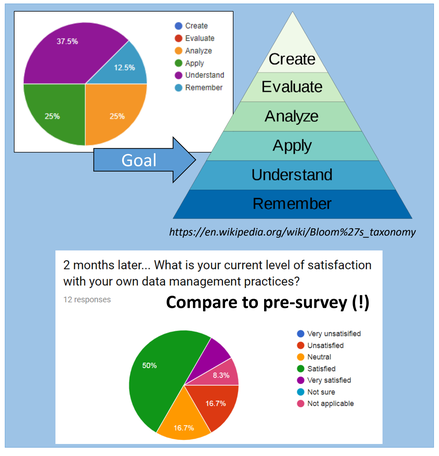
- *Yes = ~ a third of the room*

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

- *Talk to colleagues*
- *Twitter/Blogs*
- *Publications*

Post-Survey

Results from an immediate post survey indicate that participants attained greater than the intended level of learning (!).



How to reuse our materials

Our materials are available online for sharing with other oceanographic institutions, linked from our syllabus (https://tinyurl.com/WHOI-Data-Science-syllabus). Ultimately, we will publish the materials in the Woods Hole Open Access Server and provide a link from the ESIP Data Management Training (DMT) Clearinghouse. Module 1 can be used as is; perhaps modify with your own examples of challenges and solutions. For Modules 2 and 3 you would want to highlight your own institutional resources. Module 4 can be used as is, but likely you'd want to highlight contributions of scientists at your institution.

Acknowledgements

WHOI's Data Science Training Camp was developed with the support of WHOI's Academic Programs Office through a Doherty Chair in Education Award, WHOI's Ocean Informatics initiative (https://www.whoi.edu/ocean-informatics), MBLWHOI Library, WHOI Information Services (WHOI IS), and BCO-DMO (https://www.bco-dmo.org/). Helpers at the Camp included Rich Brey, Karen Soenen, Jaxine Wolfe, & Amber York. We'd also like to thank PANGEO for their BinderHub (https://binder.pangeo.io/).

References

- Bosque-Pérez, N., et al. (2016) A Pedagogical Model for Team-Based, Problem-Focused Interdisciplinary Doctoral Education. *BioScience*, 66, 477–488, https://doi.org/10.1093/biosci/biw042
- EU Marine Board (2018) Training the 21st Century Marine Professional: A new vision for marine graduate education and training programmes in Europe. Future Science Brief No. 2 April 2018, https://www.marineboard.eu/sites/marineboard.eu/files/public/publication/EMB_FSB2_2018_Web_v1.pdf
- Jordan, K., et al. (2018) Analysis of Software and Data Carpentry's Pre- and Post-Workshop Surveys. Zenodo, http://doi.org/10.5281/zenodo.1325464
- Pennington, D., et al. (2019) Bridging sustainability science, earth science, and data science through interdisciplinary education. *Sustain Sci.*, https://doi.org/10.1007/s11625-019-00735-3
- Wilson G, et al. (2017) Good enough practices in scientific computing. *PLoS Comput Biol* 13(6): e1005510, https://doi.org/10.1371/journal.pcbi.1005510

WHOI is a member of THE CARPENTRIES https://carpentries.org/

We also recommend: https://librarycarpentry.org/Top-10-FAIR/2019/01/18/oceanography/

Woods Hole Oceanographic Institution