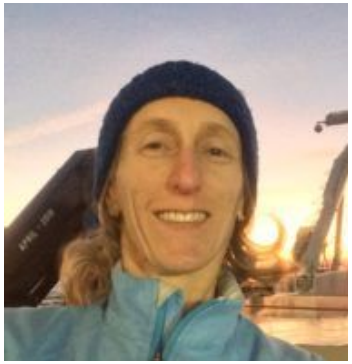
Navigating an Ocean of Data Training Camp

Sponsored by WHOI Doherty Chair in Education through Academic Programs
and co-sponsored by
WHOI Information Services, MBLWHOI Library, and WHOI Ocean Informatics initiative

January 22 & 23, 2020

Your instructors

Representing WHOI's Ocean Informatics Working Group, MBLWHOI Library, WHOI Information Services, and BCO-DMO



Stace

Lisa

Audrey

Joe

Nick

Roberta

Danie

100+ yrs > 8 hrs

And our helpers

Representing BCO-DMO, WHOI Information Services, and NES-LTER

Karen

Amber

Shannon

Rich

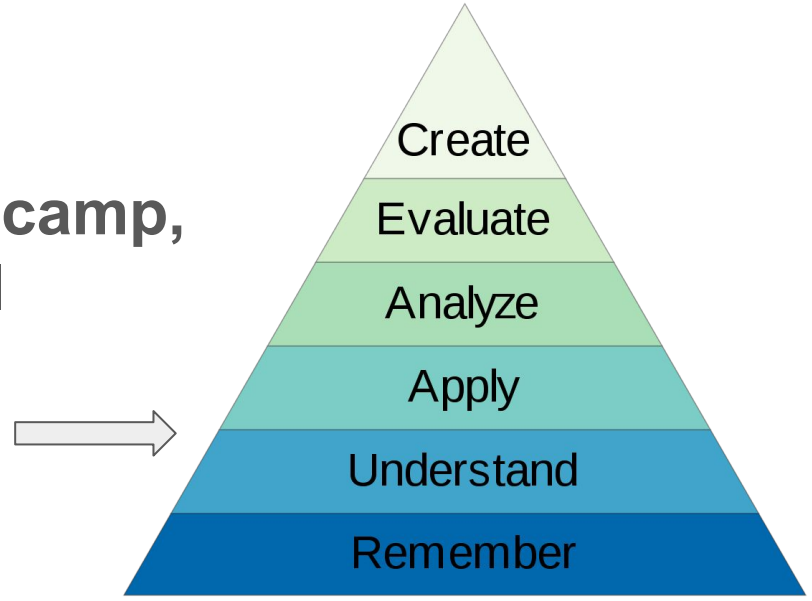
Jaxine

100+ yrs > 8 hrs

Main goals for the camp

- Foundational best practices in data, software, and project management for scientific research
- Some more specific practices in the ocean sciences
- Some resources available to you at WHOI
- Networking

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...

WHOI is a member of THE CARPENTRIES

- International organization (current members)

  "We teach foundational coding and data science skills to researchers worldwide."

- https://carpentries.org/
- Well-maintained, proven-to-work lessons
- A community, with Code of Conduct

We'll have new instructors thanks to Tech Staff Training Award in 2020!

"welcoming and supportive environment for all"

software carpentry

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

What is a “data scientist”?

This camp will use a definition within our domain *:

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

* from the Earth Science Information Partnership (ESIP) working group on 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."

Kempler, S., Mathews, T., (2017) Earth Science Data Analytics: Definitions, Techniques and Skills. *Data Science Journal*, 16(1), p.6, http://doi.org/10.5334/dsj-2017-006

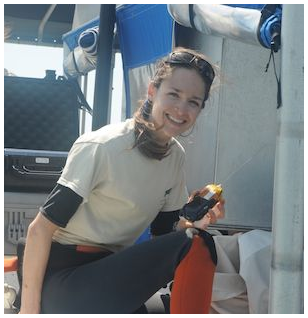
We are all data scientists

It is just to what extent

Quick online search for “data science” and JP alums:

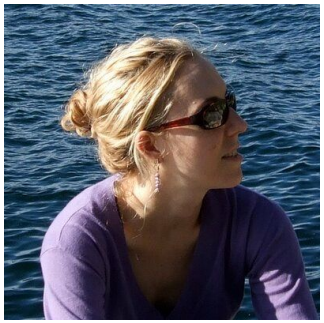
Camrin Braun
JP grad 2018

Jamie Collins
JP grad 2017



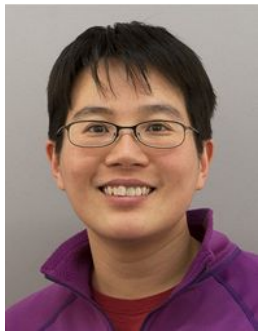
Julie van der Hoop
JP grad 2016

Harriet Alexander
JP grad 2016
Certified instructor for The Carpentries!



Carly Strasser
JP grad 2008

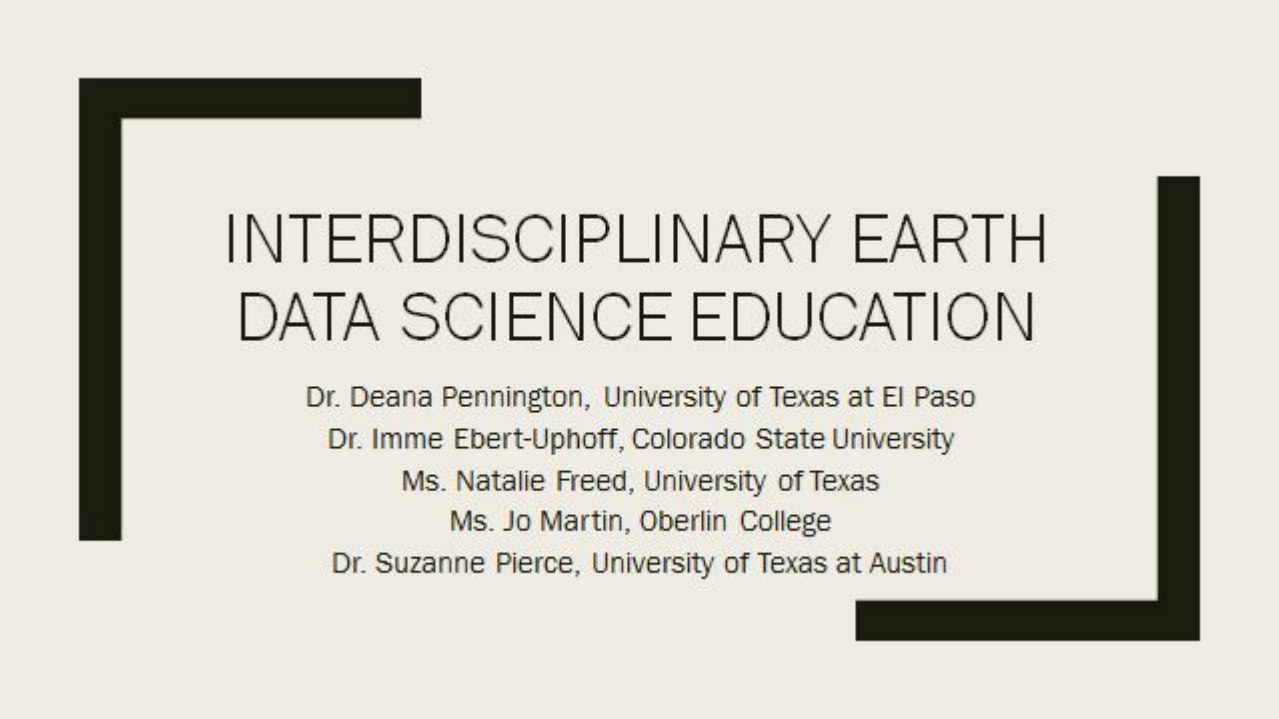
Ryan Abernathey
MIT grad 2012



Wu-Jung Lee
JP grad 2013

Ginger Armbrust
JP grad 1990

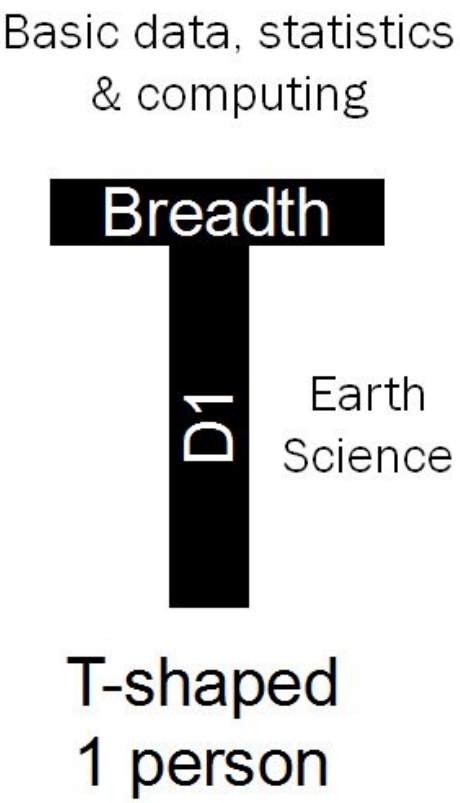
The following 4 slides adapted from:



Earth Data Science as a Topic for Interdisciplinary Education

Presentation at 2018 AGU Fall Meeting

T-shaped model



Oskam (2009)

- Engineering education: In depth knowledge in a discipline
- Broad knowledge in related fields, or in business & entrepreneurship

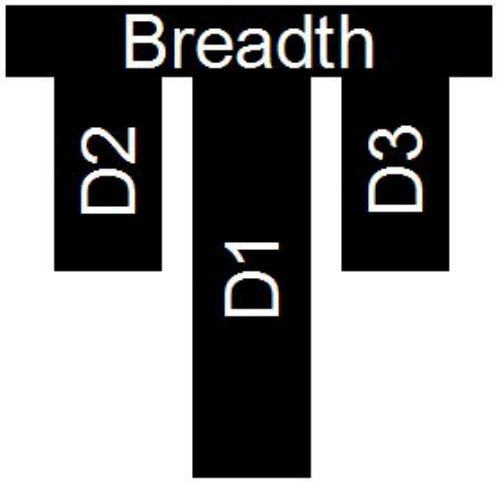
Uhlenbrook and de Jong (2012)

- Water professionals: Depth in hydrology
- Breadth in aquatic ecology and other related fields

"Training the 21st Century Marine Professional" (2018)

“Transferable Skills Training”

Shield-shaped model

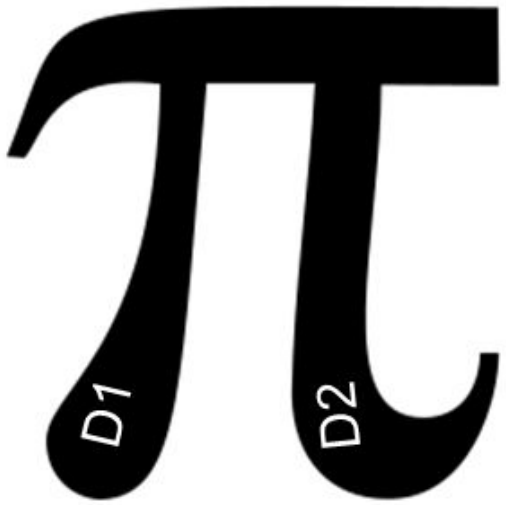


shield-shaped
1 person

Bosque-Perez et al. (2016)

Practical understanding of one or two other disciplines, sufficient to enable collaboration with researchers from those disciplines

Pi-shaped model

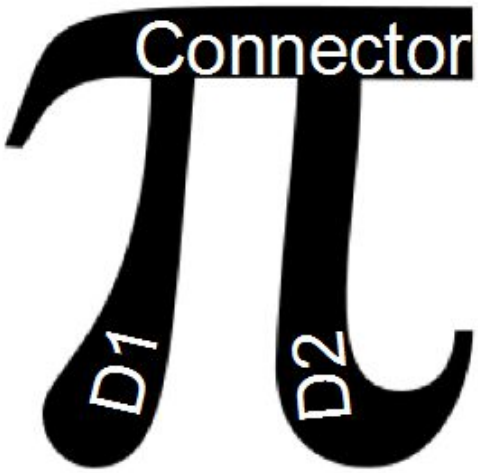


1 person

Ceri (2018) suggested a *pi*-shaped model for all science students, with *deep expertise in both a science discipline and in data science*.

However, this is not practical for many (any) PhD students, who already struggle to finish in a reasonable length of time

Pi-shaped model



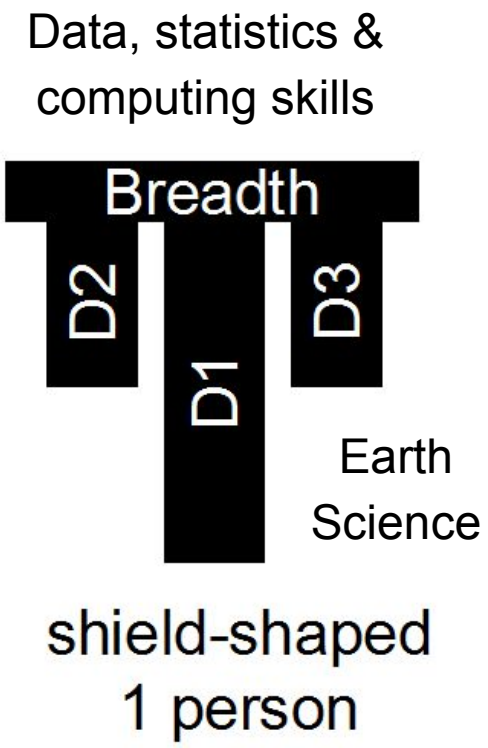
pi-shaped
2+ people

It is very unusual for a single person to reach the pi-shaped model

➔ WHOI Ocean Informatics Bridging Team

- Goal of WHOI Ocean Informatics initiative “Help WHOI staff and students adopt informatics practices and technologies”
- whoi.edu/ocean-informatics
- “What is informatics?”: Everything you do in between collecting data and what you write in your paper.

MIT's new (and huge) investment in College of Computing is aligned with shield-shape model



"The College will teach students the foundations of computing broadly and provide integrated curricula designed to satisfy the high level of interest in majors that cross computer science with other disciplines, and in learning how machine learning and data science can be applied to a variety of fields."

http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015

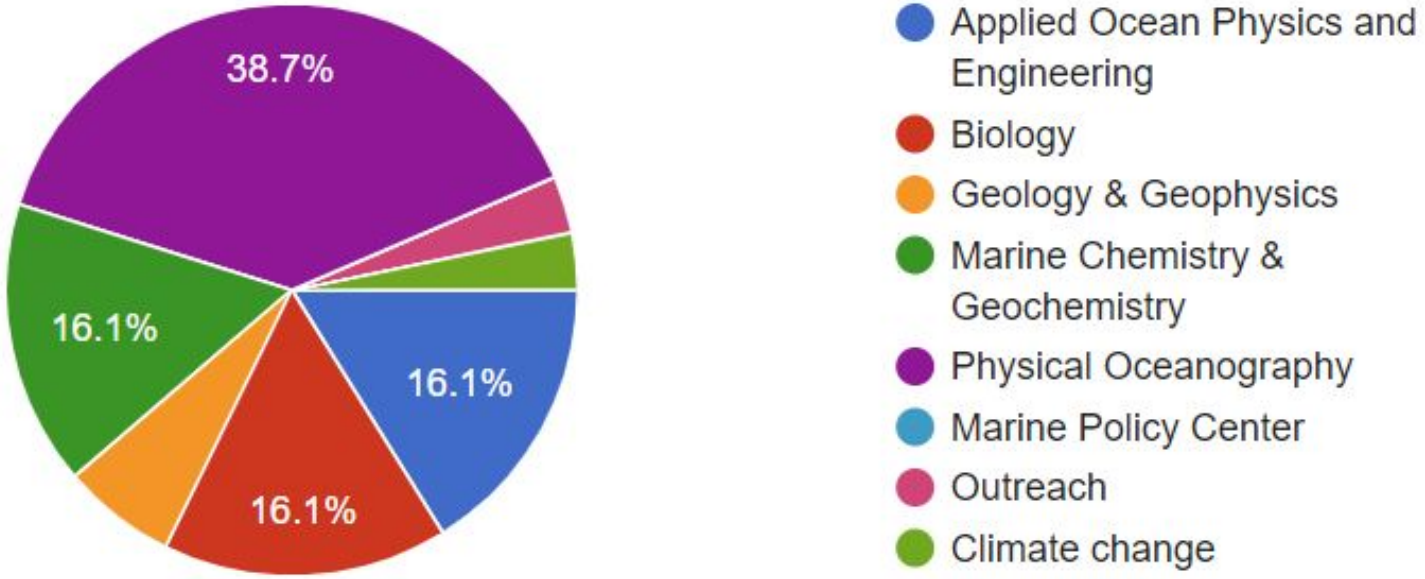
http://computing.mit.edu/about/

Why are you here for this camp?

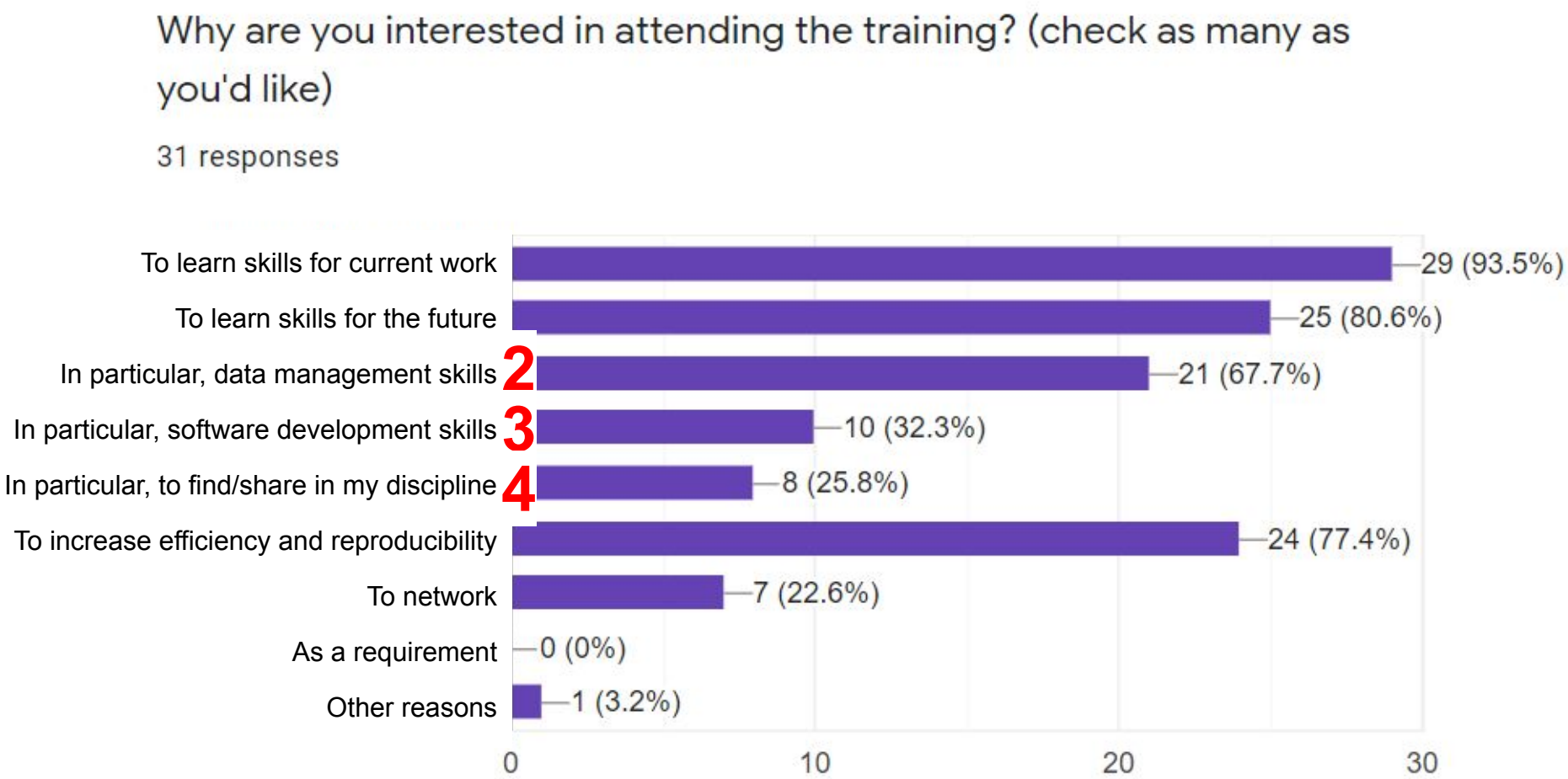
WHOI students, postdocs, and tech staff

Select which WHOI department your research aligns with (choose one)

31 responses



Why are you here for this camp?



Syllabus Module #

Link to syllabus:

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

Link to shared notes

Note that we'll post the slides from the syllabus after the workshop

12:45

Navigating an Ocean of Data Training Camp

Module 1:
"Good Enough Practices"

Main paper to guide Module 1 and set the stage for all other Modules:

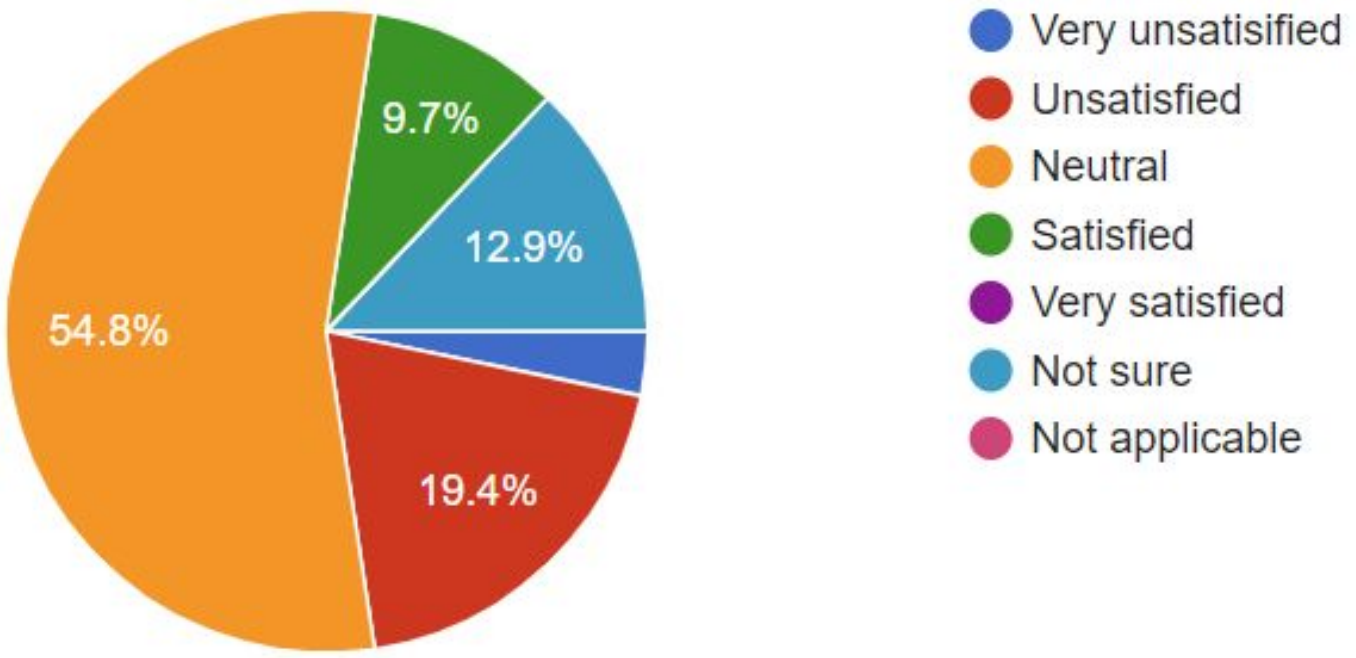
Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. *PLoS Comput Biol* 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

“Good Enough Practices”
Topics grouped for 3 discussions

1. Data Management
2. Software Development
3. Project Management and Collaboration

What is your current level of satisfaction with your own data management practices?

31 responses



“Good Enough Practices”
Data Management

a. Save the raw data.
b. Ensure that raw data are backed up in more than one location.
c. Create the data you wish to see in the world.
d. Create analysis-friendly data.
e. Record all the steps used to process data.
f. Anticipate the need to use multiple tables, and use a unique identifier for every record.
g. Submit data to a reputable DOI-issuing repository so that others can access and cite it.

“Good Enough Practices”
Data Management

a. Save the raw data.
b. Ensure that raw data are backed up in more than one location.

Challenge: How can I backup terabytes of data from research cruises?

Solution with caveats *:
WHOI Google Drive and rclone

Module 3 more options

* https://whoi-it.whoi.edu/google-drive/

Poll:

(your participation is voluntary)

How comfortable are you with managing **terabytes** of data?

- Very comfortable
- Relatively comfortable
- Neutral
- Relatively uncomfortable
- Very uncomfortable
- Not applicable

"Big Data"
V = Volume
(More V's in Module 4)

"welcoming and supportive environment for all"

“Good Enough Practices”
Data Management

c. Create the data you wish to see in the world.
d. Create analysis-friendly data.

Challenge:

- **Harmonizing spreadsheets with data contributed by multiple investigators**

Solutions:

Module 2
In-depth activity

Module 4
Domain vocabularies

“Good Enough Practices”
Data Management

e. Record all the steps used to process data.

f. Anticipate the need to use multiple tables, and use a unique identifier for every record.

g. Submit data to a reputable DOI-issuing repository so that others can access and cite it.

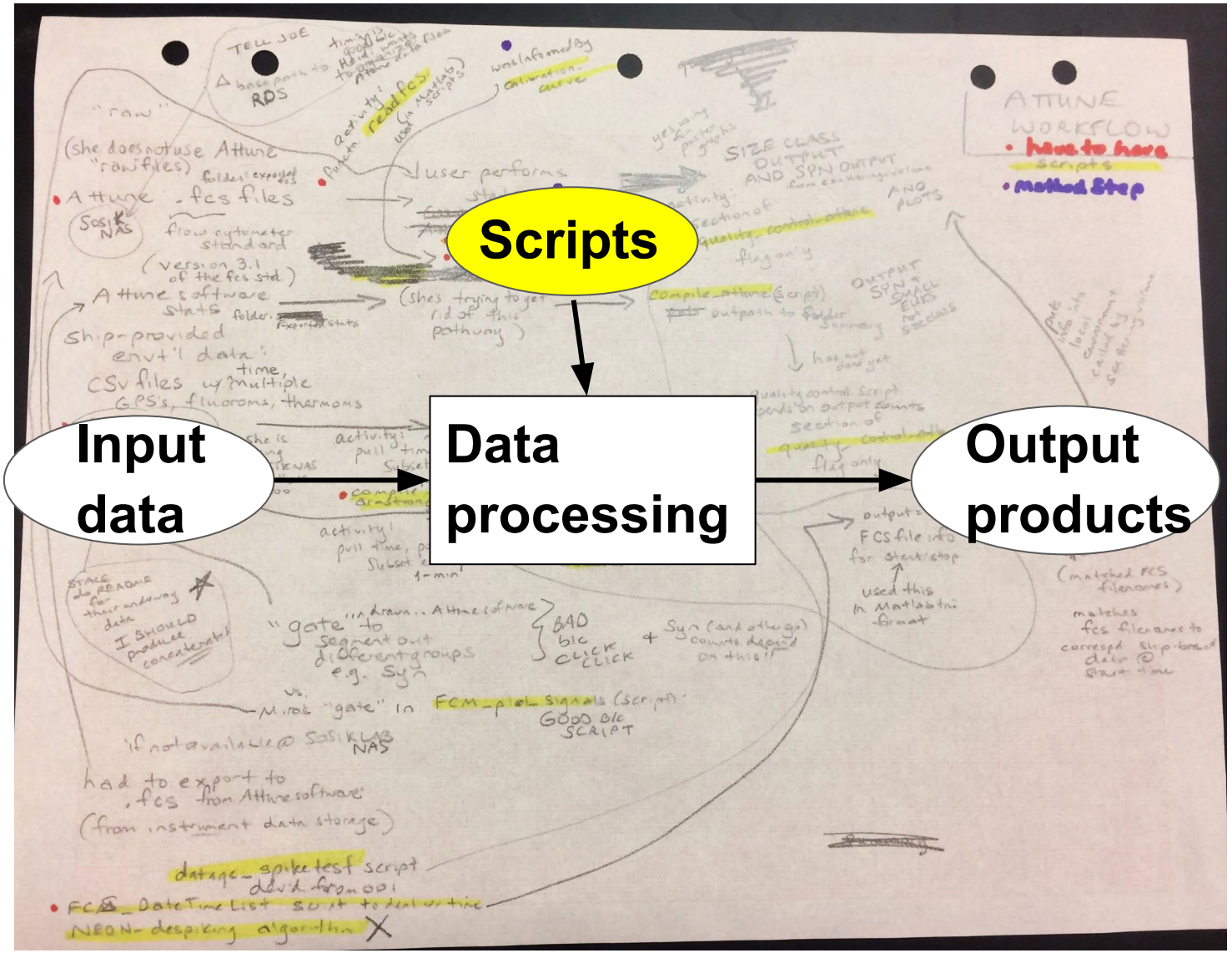
Module 4

Module 2

Solution:

Workflow diagrams

Solution: (Flowcharts and) workflow diagrams

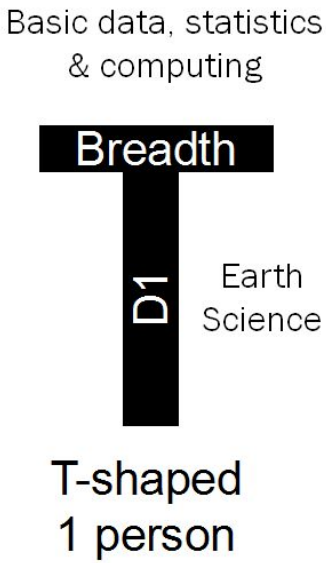


First discussion:

Data Management

Challenges and Solutions

Link to shared notes



"welcoming and supportive environment for all"

13:00 (10 min, then 5 min)

What's missing in "Good Enough Practices" Data Management?

"skills needed for effectively engaging with the heterogeneous, distributed, and rapidly growing volumes of available data…

(1) data management and **processing**,

(2) **analysis**,

(3) software skills for science,

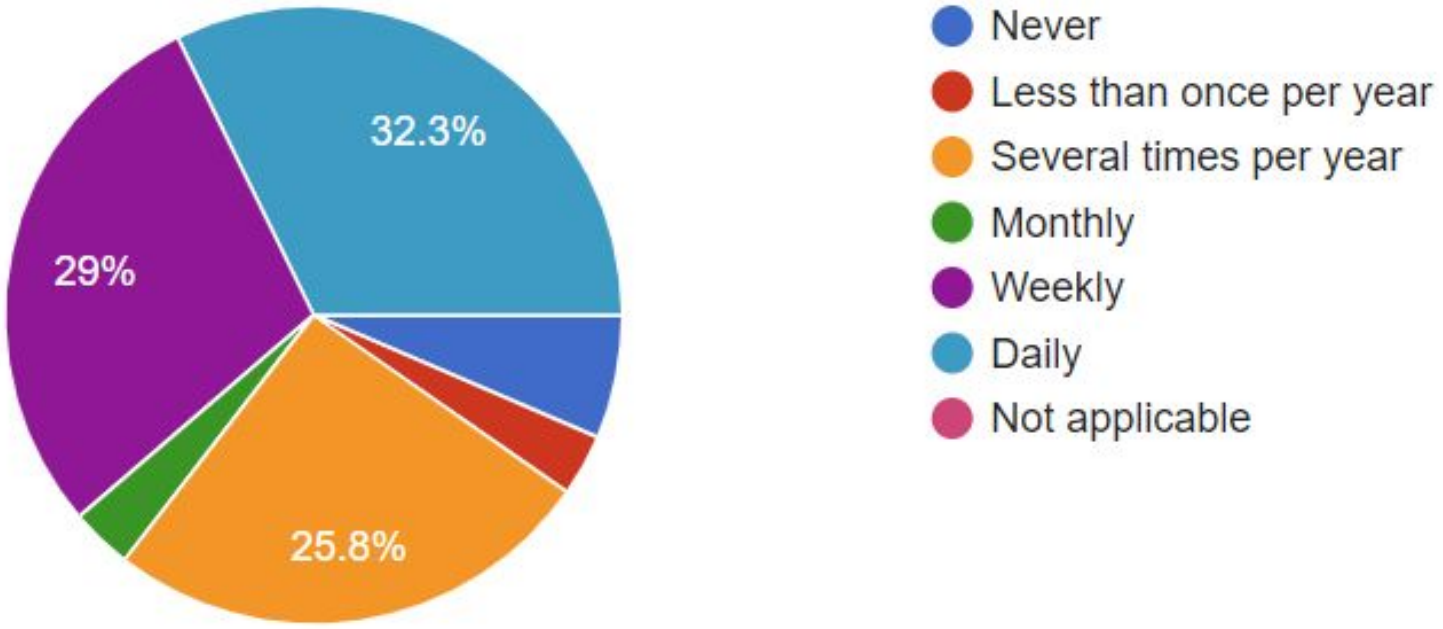
(4) **visualization**, and

(5) communication methods for collaboration and dissemination."

Hampton, S.E., et al. (2017) Skills and Knowledge for Data-Intensive Environmental Research. *BioScience*, 67, 546–557, https://doi.org/10.1093/biosci/bix025

How often do you code or program as part of your research?

31 responses



Poll

**(If you'd like, please add to shared notes.
Your participation is voluntary.)**

How do you use coding or programming in your research activities?

Poll

(your participation is voluntary)

Do you need high-performance computing - i.e., higher performance than a typical desktop computer or workstation?

Poll

(your participation is voluntary)

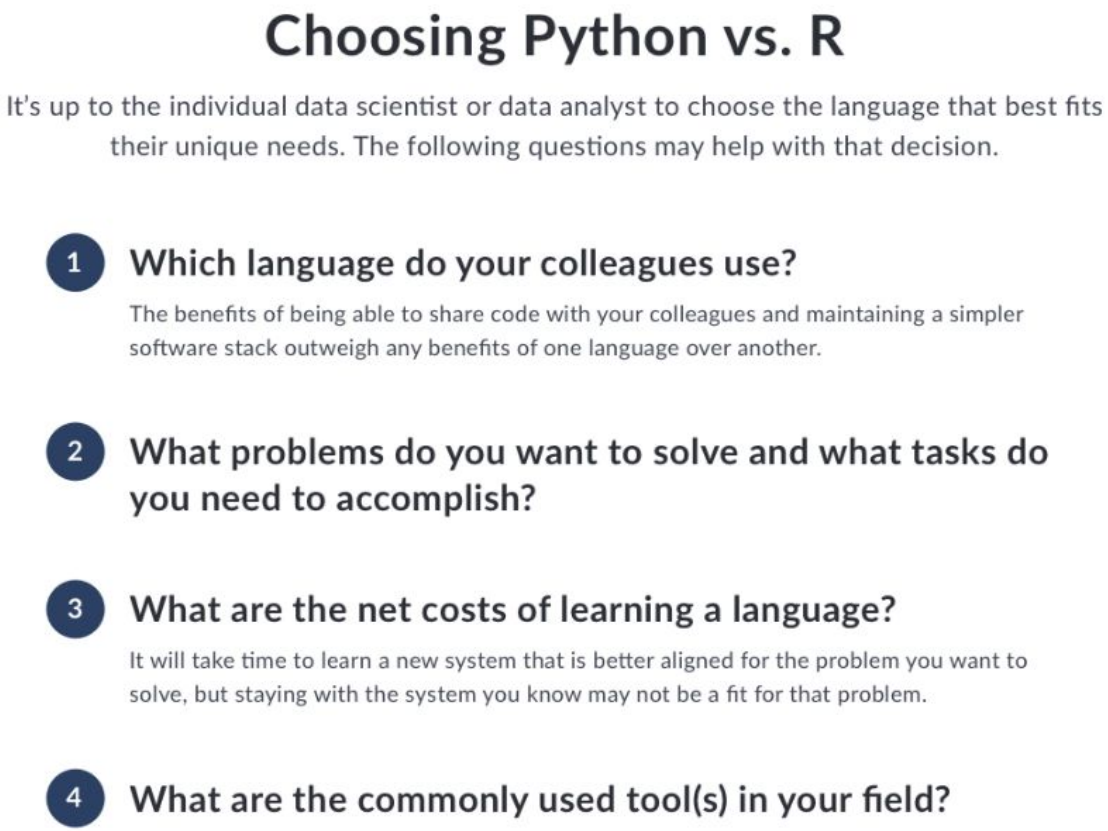
What programming language(s) do you use regularly?

- Fortran
- IDL
- Matlab
- Python
- R
- Other

Choosing Python or R?

New, just released 2 weeks ago:

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis



Choosing Python vs. R

It's up to the individual data scientist or data analyst to choose the language that best fits their unique needs. The following questions may help with that decision.

1. **Which language do your colleagues use?**

   The benefits of being able to share code with your colleagues and maintaining a simpler software stack outweigh any benefits of one language over another.

2. **What problems do you want to solve and what tasks do you need to accomplish?**

3. **What are the net costs of learning a language?**

   It will take time to learn a new system that is better aligned for the problem you want to solve, but staying with the system you know may not be a fit for that problem.

4. **What are the commonly used tool(s) in your field?**

“Good Enough Practices” Software Development

a. Place a brief explanatory comment at the start of every program.
b. Decompose programs into functions.
c. Be ruthless about eliminating duplication.
d. Always search for well-maintained software libraries that do what you need.
e. Test libraries before relying on them.
f. Give functions and variables meaningful names.
g. Make dependencies and requirements explicit.
h. Do not comment and uncomment sections of code to control a program's behavior.
i. Provide a simple example or test data set.
j. Submit code to a reputable DOI-issuing repository.

“Good Enough Practices” Software Development

a. Place a brief explanatory comment at the start of every program.

f. Give functions and variables meaningful names.

g. Make dependencies and requirements explicit.

Solution: I always prioritize these for my “future self”

“Good Enough Practices”
Software Development

b. Decompose programs into functions.

c. Be ruthless about eliminating duplication.

h. Do not comment and uncomment sections of code to control a program's behavior.

“Good Enough Practices”
Software Development

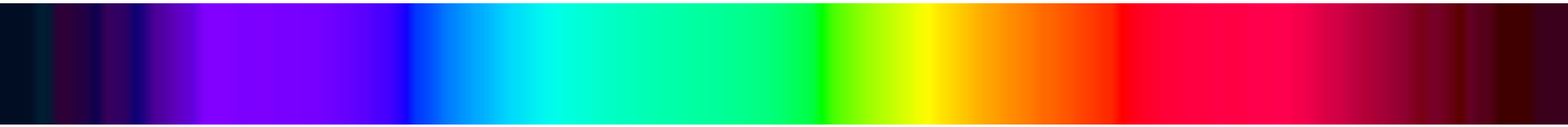
d. Always search for well-maintained software libraries that do what you need.

e. Test libraries before relying on them.

i. Provide a simple example or test data set.

j. Submit code to a reputable DOI-issuing repository.

Spectrum from script to software



Email to lab group:
“Do any of you know of efficient ways to do multi-panel bar plots (with error bars) in Matlab or R? I'd like to re-do the attached draft figure”

Challenge: how to share and give credit for code/software development in the middle of this spectrum?

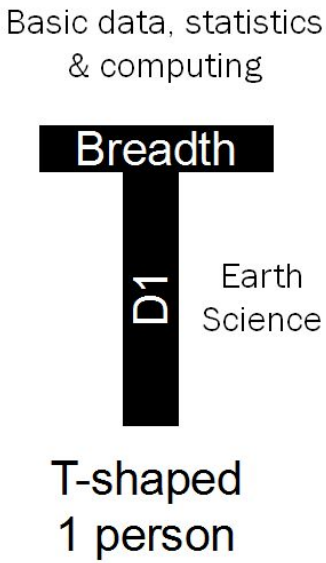
Recent JP grads who published papers for their R packages

Solution: GitHub *, and obtained unique persistent identifier from student’s institutional repository

Modules 2 & 3

2nd discussion:
Software Development
Challenges and Solutions

Link to shared notes



1:30 stretch! (5 min, then 5 min)

Data and Software and Sharing… oh my!

It's not just data, but also the metadata...

It's not just code, but the environment in which the code is run...

And then it's how you organize all of that...

And then it's how you share all of that...

This blog post is general to any data scientist and provides a nice framework for placing data and software skills into context with project management/collaboration skills:

Wheeler, S. (Feb 27, 2018) A framework for evaluating data scientist competency.

https://towardsdatascience.com/a-framework-for-evaluating-data-scientist-competency-89b5f275a6bf

“Good Enough Practices”
Project Management and Collaboration

- Project organization,
- Keeping track of changes,
- Manuscripts
- Collaboration

“Good Enough Practices”
Project Management and Collaboration

Project organization

a. **Put each project in its own directory, which is named after the project.**
b. Put text documents associated with the project in the doc directory.
c. **Put raw data and metadata in a data directory and files generated during cleanup and analysis in a results directory.**
d. Put project source code in the src directory.
e. Put external scripts or compiled programs in the bin directory.
f. **Name all files to reflect their content or function.**

“Good Enough Practices”
Project Management and Collaboration

Keeping track of changes

a. **Back up (almost) everything created by a human being as soon as it is created.**

e. Store each project in a folder that is mirrored off the researcher's working machine.

g. Copy the entire project whenever a significant change has been made.

“Good Enough Practices”
Project Management and Collaboration

Keeping track of changes

b. Keep changes small.

c. Share changes frequently.

d. Create, maintain, and use a checklist for saving and sharing changes to the project.

f. Add a file called CHANGELOG.txt to the project's docs subfolder.

h. **Use a version control system.**

Poll

(your participation is voluntary)

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

- Little or no knowledge
- Some knowledge
- Extensive knowledge

Poll

(your participation is voluntary)

Do you have a GitHub account?

Module 3: WHOlGit

“Good Enough Practices”
Project Management and Collaboration

Manuscripts

a. Write manuscripts using online tools with rich formatting, change tracking, and reference management.

b. Write the manuscript in a plain text format that permits version control.

Solutions:

- **Google doc, then if needed, move to Word with track changes**
- 

“Good Enough Practices”
Project Management and Collaboration

Collaboration

a. Create an overview of your project.

Challenge: Create a shared "to-do" list for the project.

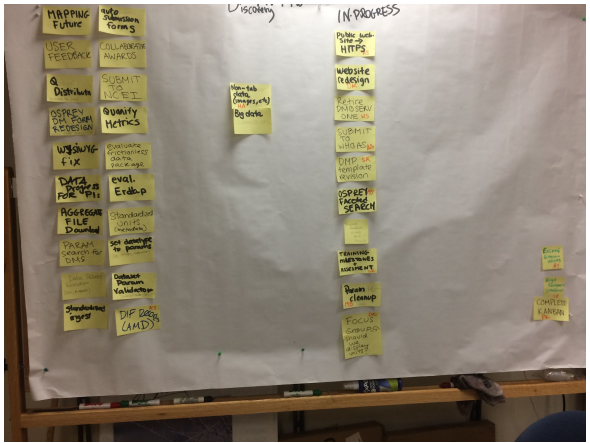
c. Decide on communication strategies.

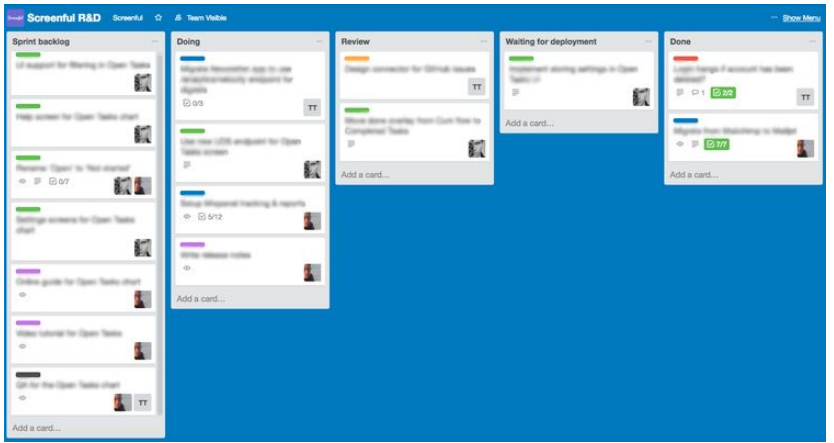
d. Make the license explicit.

e. Make the project citable.

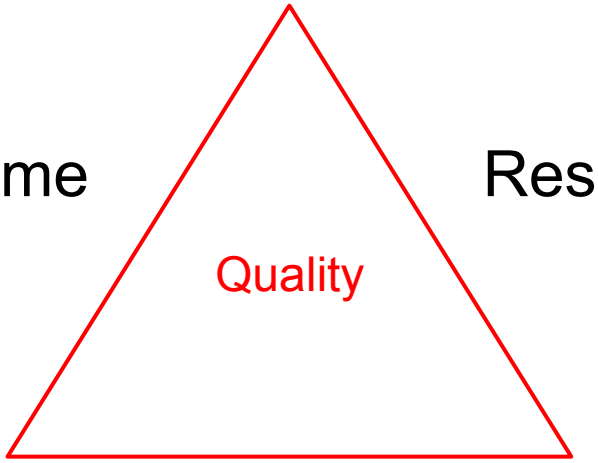
Solution:

Kanban

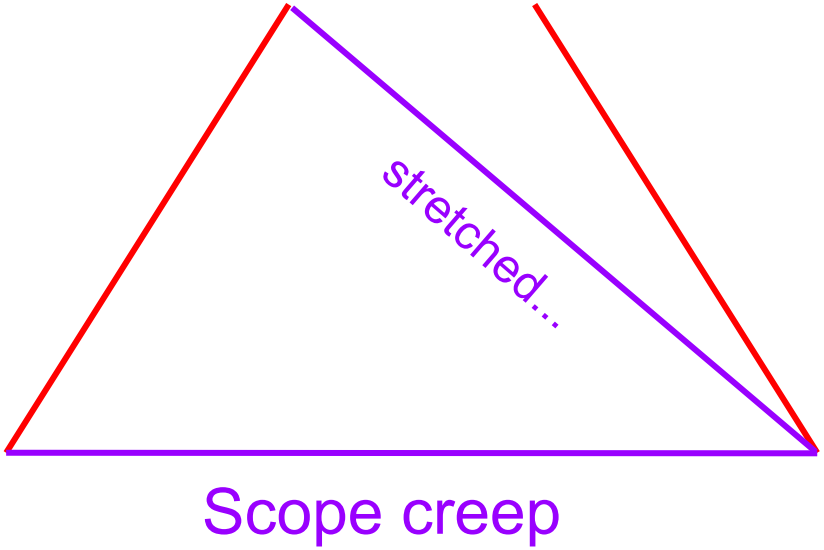




Project Management Triangle



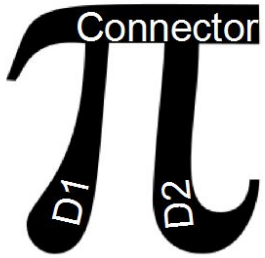
Challenge: How to avoid scope creep?



3rd discussion:
Project Management and Collaboration
Challenges and Solutions

Link to shared notes

Pi-shaped model



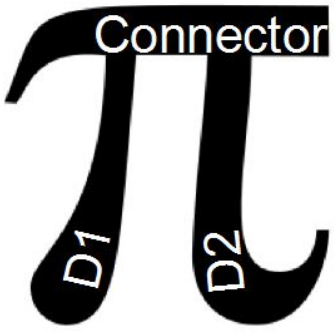
pi-shaped
2+ people

1:55 (5 min, then 5 min)

Three Steps to Successful Collaboration with Data Scientists

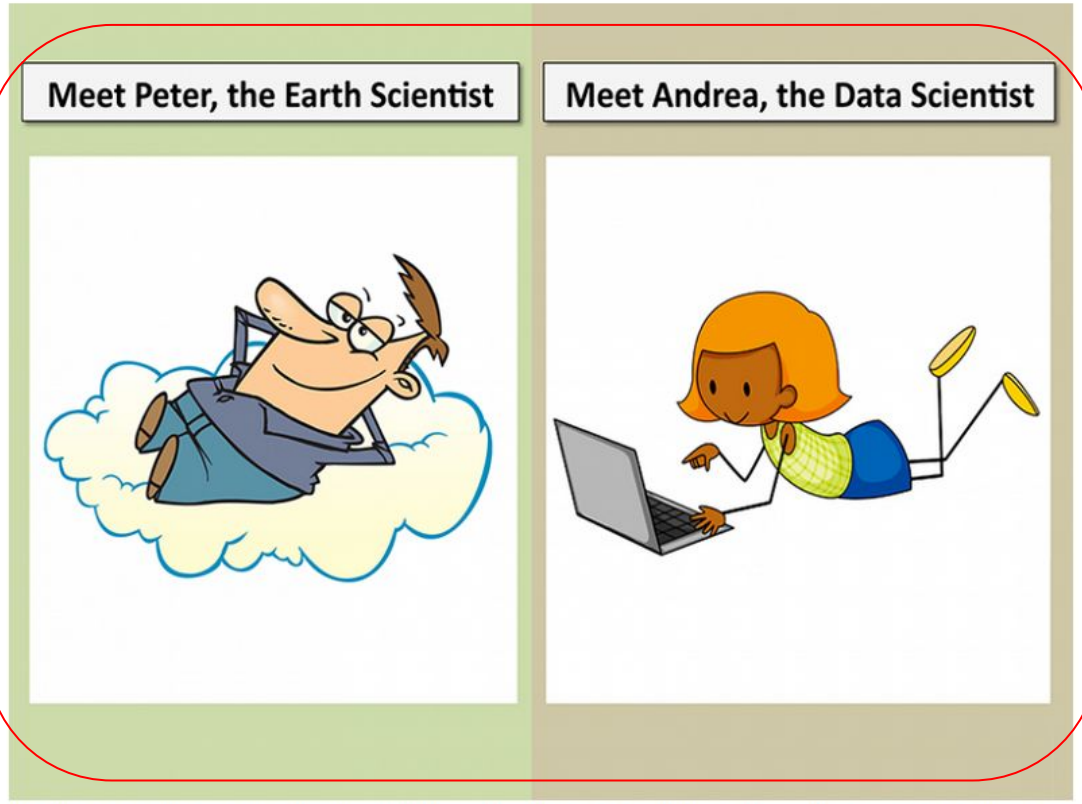
A step-by-step cartoon guide to efficient, effective collaboration between Earth scientists and data scientists.

Pi-shaped model



pi-shaped

2+ people



Peter, the Earth scientist, and Andrea, the data scientist, worked together to make important decisions regarding their research problem, approach, and experiments, as well as to validate and interpret the results of their study. Cartoon figures are from Clipart Of LLC.

Ebert-Uphoff, I., and Y. Deng (2017), Three steps to successful collaboration with data scientists, *Eos, 98,* https://doi.org/10.1029/2017EO079977

From “Good Enough Practices” Conclusion section:

“Universities can also support such efforts. While this is often provided by IT or high performance computing (HPC) groups, research librarians … have thought about and worked with data and provenance even before these computational challenges, and, increasingly, universities have dedicated data librarians on staff who have an explicit service role.”

Modules 2 & 3

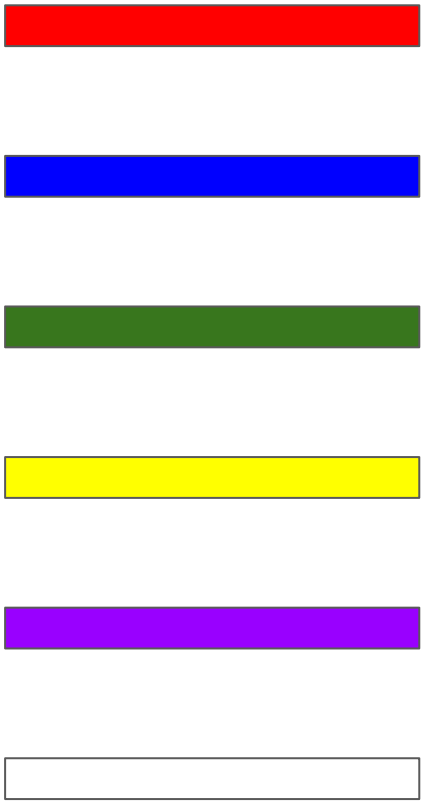
After Module 2 save the last 15 min

Describe how to prepare for tomorrow:

- Choose a journal article relevant to your research and determine if you can access the data and/or software that support the results.

Table

Please choose name tag by department.
Goal is for each table to have all depts.



AOPE

Biology

G&G

MC&G

PO

Other, or would rather not specify