Data Carpentry workshops to increase data literacy for researchers

1 Overview

The emergence of new techniques to generate or digitize data is changing the way that research is being done in many domains of research. This influx of data presents great opportunities, but also many challenges in managing, analyzing and sharing this data. However, there are not currently good training resources for researchers looking to develop these skills that will enable them to be more effective and productive researchers. To address this need, and through a supplemental grant to BEACON from the NSF, we have developed a domain-specific introductory workshop, Data Carpentry, designed to teach the basic concepts, skills, and tools for working with data.

The supplemental grant to BEACON provided resources to, among other things, extend existing computational science training material to facilitate learning by biologists and run a number of focused workshops to teach the materials and train others in delivery. Some of these resources have gone to the development and teaching of a Data Carpentry workshops, modeled after the highly successful Software Carpentry boot camps, teaching in a two-day workshop the basic concepts, skills, and tools for working with data so researchers can get more done in less time and with less pain. Enthusiasm and support for this workshop has already been overwhelming with three times as many students on the wait list as could attend at the first workshop at NESCent May 18-19. Additionally social media support, discussion and interest has been extremely positive with already dozens of requests for workshops to be run at a given location, train instructors or develop materials. Given this enthusiasm and interest from the community, we are working to continue to develop the materials, expand them to include domains beyond biology and creating a model where we can offer these workshops at locations throughout the world.

2 Data Carpentry workshops to meet data training needs

The idea for Data Carpentry came out of an NSF BIO Centers COLLABIT working group with representatives from SESYNC, NESCent, iDigBio, BEACON, NEON and iPlant. There was a shared need across the BIO Centers for increased researcher knowledge in data management and analysis and computational literacy.

Dr. Tracy Teal from MSU, funded on the BEACON supplement, led a breakout section to talk about developing a 'data and computational literacy' workshop. Many of the attendees had taken or taught Software Carpentry workshops and had seen the success of the model for materials development and training. After the workshop we completed surveys across the BIO centers to determine the needs of users. Using this information, we created learning objectives and designed the workshop to teach the following topics.

Objectives for learning: Researchers should be able to retrieve, view, manipulate, analyze and store their's and other's data in an open and reproducible way.

Course topics

- How to use spreadsheets (Excel) more effectively and the limitations of spreadsheets
- Getting data out of Excel and in to more powerful tools - R or Python
- Introduction to databases, managing and querying data in SQL
- Workflows and automating repetitive tasks with an introduction to the shell

Included discussions

- Preparing data for analysis
- Using data or computational resources - publicly available databases and resources for computation, such as Amazon or iPlant Atmosphere
- An emphasis throughout the workshop on reproducible and open research

While many of the topics in Data Carpentry are similar to that of Software Carpentry, the Data Carpentry workshop differs in its focus, its level of expected knowledge and its domain specificity.

- *Data Carpentry is focused on data.* The workshop introduces one data set at the beginning of the workshop. Throughout the workshop, it is taught how to manage and analyze that data in an effective and reproducible way.
- *Data Carpentry is designed for novices.* There are no prerequisites and no prior knowledge about the tools is assumed.
- *Data Carpentry is designed to be domain specific.* Researchers learn better when the example used is one that they are familiar with. The information is more easily integrated in to an existing framework and learners are more motivated by examples like their own work.

3 Overview and interest in the first Data Carpentry workshop

The first Data Carpentry workshop was taught at NESCent May 18-19th with instructors who developed and taught the material from NESCent (Dr. Hilmar Lapp and Dr. Karen Cranston), BEACON (Dr. Tracy Teal) and the Utah State Univ. (Dr. Ethan White) and teaching assistants from SESYNC (Dr. Mike Smorul) and iDigBio (Dr. Deborah Paul, Darren Boss and Matt Collins) with travel support for instructors coming from the NSF BEACON supplement, DataONE and iDigBio.
http://nescent.github.io/2014-05-08-datacarpentry/

There were 30 spots available for the course, and it was full within 3 hours of it being announced. 64 people signed up on the wait list, and anecdotal evidence suggests that many people who were interested didn't even sign up for the wait list. This immediate interest highlighted the need for this type of course.

In the introductory section where learners discussed why they were taking the course, many reasons were mentioned, including frustration with current data management and analysis approaches, an interest in advancing their research, teaching the tools to others and in learning the skills for future career goals. Some sample comments included:

- I'm tired of feeling out of my depth on computation and want to increase my confidence.
- I usually manage data in Excel and it's terrible and I want to do it better.

- I'm trying to reboot my lab's workflow to manage data and analysis in a more sustainable way.
- I want to use public data.
- I work with faculty at undergrad institutions and want to teach data practices, but I need to learn it myself first.
- I'm interested in going in to industry and companies are asking for data analysis experience.
- I'm re-entering data over and over again by hand and know there's a better way.
- I have overwhelming amounts of NGS data.

This first course taught the topics discussed above, and we learned that there is a need for refinement and a re-ordering of topics. Overall the workshop was well received however with positive comments to instructors and on social media. Post-assessment ratings gave the course an average rating of 8.25 out of 10.

A write-up of that workshop is available on the Software Carpentry blog
http://software-carpentry.org/blog/2014/05/our-first-data-carpentry-workshop.html

Interest beyond direct workshop participants has come from email to instructors, over Twitter and in response to blog posts from researchers in biology, genomics, digital humanities, library sciences and social sciences. This response has included an interest in hosting future workshops, learning to be instructors so people could teach it locally, interest in helping to develop materials and in taking the course themselves. Interestingly, there was a lot of interest from librarians and people working at university libraries. With recent data management requirements from the NSF and other funding agencies, university libraries have taken on the challenge of helping researchers develop data management plans, track data provenance and distribute and share data. Many libraries provide multiple resources and are actively developing or running workshops on these topics, so Data Carpentry workshops seem to fit it well with their interest and engagement model.

Additionally there is interest from the UK based ELIXIR whose goal it is to "build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society" in being involved in the development and distribution of this course.

4 Next steps for Data Carpentry

We would like to continue to train researchers in good data analysis and management practices, building off the enthusiasm and momentum of Data Carpentry and move forward with more workshops and organization to meet these goals.

4.1 Initial next steps

The initial next steps for Data Carpentry are to run more workshops at the NSF BIO Centers. This will help to iterate through the materials, so they can be refined and the course adjusted to better meet learner needs. Many of the same instructors for the NESCent workshop are continuing to work to refine and develop materials and workflow.

A workshop is scheduled at BEACON for July 24-25 and another is tentatively scheduled for July 17-18. Karen Cranston, Hilmar Lapp, Ethan White, Tracy Teal and Deborah Paul continue to be involved in the course development. Teal from BEACON will be an instructor at both workshops and is training existing Software Carpentry instructors at BEACON on teaching the Data Carpentry materials. Many of those Software Carpentry instructors were trained at a recent train-the-trainer workshop also supported by the BEACON supplemental funding.

4.2 What would success for Data Carpentry look like

Goal: Develop and teach Data Carpentry workshops to help train the next generation of researchers in good data analysis and management practices to enable individual research progress and open and reproducible research.

The challenge is to meet this need for data training in many different locations and across multiple domains of interest. If a researcher wants to take a Data Carpentry course, our hope is that we will be able to provide that resource.

Towards that aim there are several longer term goals:

- The ability to host workshops in many different locations, even to run them multiple times in the same location.
- Materials developed for domains of interest
- Materials in both R and Python
- A streamlined assessment where we can assess learning in a workshop and its effects as researchers progress through their careers
- A forum for continued engagement on Data Carpentry post workshop
- A set of resources where learners could look for more information on particular topics (Software Carpentry already has good online tutorials for many of these topics and there are other good exiting online resources)
- More advanced workshops - data visualization, more advanced R or Python for statistics

How these needs can be met.

Already in place:

- github for the development and distribution of materials
- online SWC tutorials on many topics
- first set of materials developed by SWC and adapted for NESCent workshop by personnel supported by NSF funding
- some support from SWC, the Sloan Foundation and the Mozilla Science Labs for the logistical organization of workshops

To be established:

- Personnel for establishing workshop guidelines and structure
- Personnel for materials development and coordinating efforts
- Train the trainer workshops to train many instructors so that they can host local workshops
- The development of assessment materials to assess learning in workshops and longer term impacts on researcher's data practices

One idea is to run Data Carpentry similar to a franchise, with a strong train the trainers component that would allow instructors to run many workshops locally, reducing coordination efforts and travel costs for instructors. Many people feel qualified to teach the materials covered in Data Carpentry with some training on how the workshop should be taught. This differs somewhat from Software Carpentry, where the topics are perceived to be a bit more technical and local personell have less confidence in their ability to teach the materials.