Data Carpentry:
workshops to increase data literacy for researchers

Tracy K. Teal* Karen A. Cranston† Hilmar Lapp‡ Ethan White§

June 10, 2014

1 Overview

In many domains of science the rapid generation of large amounts of data is fundamentally changing how research is done. The deluge of data presents great opportunities, but also many challenges in managing, analyzing and sharing data. Good training resources for researchers looking to develop skills that will enable them to be more effective and productive researchers are scarce. To address this need we have developed an introductory bootcamp-style workshop, Data Carpentry, designed to teach basic concepts, skills, and tools for working more effectively with data.

Using the highly successful Software Carpentry bootcamps as a model, we developed Data Carpentry as a two-day workshop that teaches basic concepts, skills, and tools for working with data so researchers can get more done in less time and with less pain. We modified existing novice training materials from Software Carpentry, developed new lessons, and integrated them in to a workshop focused on data and designed to facilitate learning by researchers with little to no prior knowledge of programming, shell scripting, and command line tools.

The first workshop was held at the National Evolutionary Synthesis Center (NESCent) on May 8-9, 2014. Enthusiasm and support has already been overwhelming. The 30 seats available for registration were gone in less than 24 hours, and three times more students attempted to register than could be accommodated. Social media support, discussion and interest has been extremely positive with dozens of requests for workshops to be run, instructor training, and material development. Based on the enthusiasm and interest from the community, we are eager to continue developing materials, to expand them to include domains beyond biology, and to create a model scalable enough so we can offer these workshops at locations throughout the world.

*Michigan State University, East Lansing, MI, USA
†National Evolutionary Synthesis Center (NESCent), Durham, NC, USA
‡National Evolutionary Synthesis Center (NESCent), Durham, NC, USA
§Utah State University, Logan, UT, USA

2 Data Carpentry workshops to meet data training needs

Many organizations and groups have been working to develop a Data Carpentry course, including ELIXIR-UK[1], ANDS[2], and BIO CollabIT[3]) due to the shared need for better training researchers in advanced data management, analysis, and computational literacy. At a BIO CollabIT meeting in September 2013, informatics staff from several of the represented interdisicplinary science centers developed the cornerstones of a data and computational literacy workshop based on the Software Carpentry model. Many of the attendees had taken or taught Software Carpentry workshops, and had seen firsthand the success of this model.

Based on our experiences supporting researchers engaged in interdisciplinary collaborative science, and informed by surveys conducted across the NSF BIO-funced centers, we defined the following overall learning objective for the course:

> ***Learning objective:*** *Researchers should be able to retrieve, view, manipulate, analyze and store their and other's data in an open and reproducible way.*

To attain this objective, we identified the following teaching subjects.

- How to use spreadsheet programs (such as Excel) more effectively, and the limitations of such programs.
- Getting data out of Excel and into more powerful tools — using R or Python.
- Using databases, including managing and querying data in SQL.
- Workflows and automating repetitive tasks, in particular using the command line shell and shell scripts.

In addition to the above subjects, the following skills emerged as particularly important to impart from our discussions about designing the course:

- Preparing data for analysis.
- Using data and computational resources, in particular publicly available ones such as Amazon Web Services or iPlant Atmosphere.
- Conducting data and computation-heavy research more reproducibly and openly.

Although the topics for Data Carpentry overlap substantially with those for 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. This data set is used throughout the workshop to teach how to manage and analyze 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 domain specific by design.* Researchers learn better when the example used is of a kind they are familiar with. Learners can more easily integrate new skills and information into an existing framework, and are more motivated by example data similar to data encountered in their own work.

[1] `http://elixir-uk.org/`

[2] Australian National Data Service; `http://www.ands.org.au/`

[3] An informal consortium of science-supporting IT groups at interdisciplinary centers funded by the NSF BIO directorate; `http://www.nescent.org/wg_collabsci/`

3 Overview and interest in the first Data Carpentry workshop

The first Data Carpentry workshop took place at NESCent May 8-9, 2014[4]. Instructors who developed and taught the material were from NESCent (HL and KAC), BEACON (TKT) and Utah State University (EW). Assistants, most of whom are planning to run and teach Data Carpentry courses themselves in the near future, were from SESYNC[5] (Dr. Mike Smorul), iDigBio[6] (Deborah Paul and Matt Collins) and iPlant[7] (Darren Boss).

The course provided room for 30 learners, and it was full within 3 hours of opening registration. An additional 64 people registered for the wait list, and anecdotal evidence suggests that there were many people who were interested but did not register for the wait list once they saw that the course was full. This immediate interest demonstrates the degree of need for this type of course.

The workshop started with learners describing why they were taking the course. The reasons included frustration with current data management and analysis approaches; an interest in advancing their research; teaching the tools to others; and learning the skills required for future career goals. The following examples illustrate the range of motivations:

- *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 I know there's a better way.*
- *I have overwhelming amounts of next-generation sequencing data.*

This first course taught the topics outlined in Section **??**. Overall the workshop was well received, with positive comments to instructors and on social media. Post-assessment ratings gave the course an average rating of 8.25 out of 10. The experience of teaching the material and feedback from learners also gave rise to ways of refining both the topics taught and the order in which they are taught, as discussed in a post-workshop write-up[8].

Broader interest in the workshop has been expressed through email to instructors, over Twitter and in blog post responses from researchers in biology, digital humanities, library sciences, and social sciences. Interests have ranged from hosting future workshops, to teaching workshops and helping to develop materials.

Librarians and people working at university libraries have been particularly interested in Data Carpentry. With recent data management requirements from the NSF, NIH 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 fit well with their interest and engagement model.

[4]Course website at `http://nescent.github.io/2014-05-08-datacarpentry/`

[5]`http://sesync.org`

[6]`http://idigbio.org`

[7]`http://iplantcollaborative.org`

[8]`http://software-carpentry.org/blog/2014/05/our-first-data-carpentry-workshop.html`

This workshop also aligns well with training activity proposals labeled "Data Carpentry" from ELIXIR-UK. ELIXIR's goal is to "build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society", and there is potential for mutual engagement between ELIXIR-UK and our group on the further development of this course.

4 Next steps for Data Carpentry

Building off the enthusiasm and momentum of Data Carpentry, we want to continue training researchers in good data analysis and management practices and move forward with more workshops and organizational support for instructors and materials development to meet these goals.

4.1 Initial next steps

As the initial design of the workshop resulted from data management training gaps identified as common among the NSF BIO-funded science centers, running more workshops at other centers is the logical next step for Data Carpentry. This will help further polish the materials so that the course meets learners' needs at different centers, in different fields of (biological) science. Many of the instructors and assistants for the NESCent workshop have continued to refine and develop the materials and workflow.

The next such workshop is already scheduled at BEACON for July 24-25. TKT (BEACON) will be an instructor at the upcoming workshop and is also training existing Software Carpentry instructors on teaching the Data Carpentry materials.

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, including running 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
- More advanced workshops - data visualization, more advanced R or Python for statistics

What will be needed to achieve these goals?

Several components are already in place or under active development:

- Github repositories for the development and distribution of materials
- online Software Carpentry tutorials on many topics
- First set of materials adapted from Software Carpentry for the Data Carpentry workshop at NESCent
- Some support from Software Carpentry and the Mozilla Science Labs for the logistical organization of workshops

The components and capabilities yet to be established include the following:

- Personnel for establishing workshop guidelines and structure
- Personnel for materials development and coordinating efforts
- Train-the-Trainers workshops to expand the group of instructors, to increase the number of institutions with local trainers, and to decrease the requirement for instructor travel
- The development of assessment materials and mechanisms to measure learning in workshops and to assess longer term impacts on researcher's data practices

One idea is to operate Data Carpentry similar to a franchise, with a strong Train-the-Trainers component that would enable instructors to run workshops under the Data Carpentry brand at their local institutions, with central logistical support for registration, coordination of material development, and other tasks that benefit substantially from economies of scale. Aiming for instructors to be typically local to where a course is run also reduces instructor coordination and travel costs. People who expressed interest in teaching Data Carpentry tend to self-report as qualified to do so, provided they can receive some training on didactics and pedagogy. This is in contrast to what we typically observe for Software Carpentry, suggesting that Data Carpentry is particularly well-suited to the franchise-model of scaling up.

5 Acknowledgements

The NESCent workshop could not have been developed or taught without the support of several organizations, including for personnel time, materials, and travel funds. Specifically, DataONE (NSF #083094) supported the work of HL and KC on this workshop; NSF (NSF Career award) supported EW; BEACON (NSF) supported TKT. Assistants: Mike Smorul was supported by SESYNC (NSF DBI-1052875), Deb Paul and Matt Collins by iDigBio (NSF) and Darren Boss by iPlant (NSF) for travel and work on materials development.

We are also grateful to Software Carpentry and the Mozilla Science Lab for guidance on materials development and administrative support.