IDCC15 | *Practice Paper*

Data Carpentry: workshops to increase data literacy for researchers

Tracy K. Teal
Michigan State University, East Lansing, MI, USA

Karen A. Cranston
National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

Hilmar Lapp
National Evolutionary Synthesis Center (NESCent), Durham, NC, USA

Ethan White
Utah State University, Logan, UT, USA

Greg Wilson
Software Carpentry Foundation, Toronto, Canada

Aleksandra Pawlik
University of Manchester, United Kingdom

Abstract

In many domains 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 two-day intensive workshop, Data Carpentry, designed to teach basic concepts, skills, and tools for working more effectively and reproducibly with data.

A survey of researchers in the National Science Foundation's BIO Centers revealed not only gaps in knowledge in data management and analysis, but also a frustration by researchers in their inability to or struggle with skills in the data lifecycle. Researchers felt limited by spreadsheets, were doing time consuming tasks by hand, wanted to use public datasets, had next-generation sequencing datasets they were unable to even open and were finding external hard drives with theirs or colleaugues untracked datasets. Fundamentally this lack of skills and of confidence is limiting research progress. Researchers know they wanted to learn more and increase their 'data literacy' but were unsure of where to get the training. This is a sentiment echoed generally across now many domains of research.

Draft from 13th October 2014

Correspondence should be addressed to Aleksandra Pawlik, Room 1.17 Kilburn Building, Oxford Road, University of Manchester, M13 9PL, Manchester, United Kingdom. Email: aleksandra.pawlik@manchester.ac.uk

The 10th International Digital Curation Conference takes place on 9–12 February 2014 in London. URL: http://www.dcc.ac.uk/events/idcc15/

(Citations for need for data skills and lack of training opportunities would be nice - TKT will look for ones she has) Many organizations including ELIXIR-UK and the Australian National Data Service have identified this need as well.

Software Carpentry, two-day hands-on bootcamp style workshops teaching best practices in software development, have demonstrated the success of short workshops to teach foundational research skills. We have adpated this model with the objective of teaching skills to researchers to enable them 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 spreadsheets 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.

An alternative way to describe the workshops:

- Workshops should be domain specific. Each field has its own data types, analysis packages and standard problems to address. Being able to teach people in their domain letâĂŹs them both more immediately understand the questions and approaches, but then be able to apply it to their own work. Using examples that are 'real world' to a given domain is fundamentally motivating for the skills that are being taught.

- Workshops should be a narrative that show the data lifecycle for a given dataset or problem. All components of the data lifecycle are fundamental in the quality of the final analysis. Emphasizing all the components from setting up data tables, to viewing, manipulating, analyzing. visualizing and sharing data is crucial for accurate outcomes and reproducible research. Also, this lifecycle again models a users' workflow, allowing learners to put the process in to action with their own data sets.
- Workshops are designed for people with no prior computational experience. Learners can walk in with any level of background, but these workshops assume no prior knowledge. In this way learners should not self-select whether or not they should attend, and there is clear expectation for the pace of instruction. We also can meet researchers where they are and build on existing practices and knowledge.
- These workshops can be focused on any research domain, not just science. Social scientists, digital humanists, librarians, and museum collections are also facing the same challenges with the digital data deluge. The same principles in the data lifecycle can be applied in any domain of research and materials adapted to meet the specific needs of that domain.
- Running an effective workshop means having instructors trained in how to teach, particularly in a workshop format. SWC has developed an effective train-the-trainers program based on pedagogy and experience. Workshops should be taught by at least one SWC trained instructor.

Four Data Carpentry workshops in biology have now been taught with positive response and survey assessment results demonstrating that learning objectives are being met. The research community has been enthusiastic about hosting, teaching or taking these workshops as well as been engaged in the delveopment of materials in other domains and in expanding topics. Work is already progressing on materials for genomics, neuroscience, social science and geoscience and is expanding to include lessons on data visualization and introductory statistics.

Data Carpentry workshops won't be able to teach researchers all the skills they need in two days, but we've shown that they are a way to get the process started and that they can lower the activation energy required for researchers to be able to do more and more effective work with their data and enable research progress.