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

Karthik Ram
Section of Evolution and Ecology, University of California, Davis, CA, USA

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. However good training resources for researchers looking to develop skills that will enable them to be more effective and productive researchers are scarce and there is little space in the existing curriculum for courses or additional lectures. 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.

These workshops are based on Software Carpentry, two-day hands-on bootcamp style workshops teaching best practices in software development, that have demonstrated the success of short workshops to teach foundational research skills. Data Carpentry focuses on data literacy in particular, 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 in order to extract knowledge from data.

Draft from 13th March 2015

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 [TBC] in [TBC]. URL: http://www.dcc.ac.uk/events/idcc15/

Introduction

With the increasing ability to digitize text and collections, automate data collection, conduct large scale surveys and generate vast genomic, geophysical or other type of data, there is the great potential to conduct data-driven research and address questions in all fields that were not previously possible. However, despite this promise, analysis of these datasets presents a major challenge. Many researchers lack the computational and statistical training required for appropriate data analysis or a working vocabulary to communicate their analysis needs to computer scientists or statisticians. Many researchers are unfamiliar with best practices and tools in the data lifecycle: most or all of what they know about data management, analysis, and sharing has been learned piecemeal, or not learned at all.

This is especially concerning because it limits the ability of researchers to make progress on these important questions or results in inaccurate or incomplete analyses that can lead to erroneous conclusions. It also leads to the generation of data that ultimately cannot be used to answer research questions, wasting or underutilizing available resources.

However, many researchers are interested in developing better approaches to how they manage and analyze data. A survey of researchers in the National Science Foundation's BIO Centers revealed some gaps in knowledge in data management and analysis, but also that researchers are frustrated with their current data workflows and know their research capacity is being limited by this lack of knowledge. In a Bioinformatics Resource Australia EMBL 2013 Community Survey Report the most emphatic outcome was the overwhelming demand for training[1]. More than 60% of researchers surveyed said that their greatest need was additional training, compared to a meagre 5% who need access to additional compute power. While this survey is focused on biology and bioinformatics, the sentiment is shared by researchers in many domains and regions and has been clearly identified by the ELIXIR-UK[2] project as well. Fundamentally this lack of skills and of confidence is limiting research progress.

However, good training resources for researchers looking to develop these skills are scarce. and it is difficult to determine where to start. Training in data and computing skills is still largely absent from undergraduate and graduate programs. Self-guided study such as online lessons, MOOCs and books are available, but there is a significant challenge in being able to discover relevant and high-quality materials and for already busy researchers to commit their time and focus to these learning activities. The completion rate for MOOCs in particular is less than 10%[3]. Also training resources are often available in areas outside a resarcher's domain, so they have the added challenge of figuring out how to apply learned tools or approaches in the context of their research. Instead, most researchers learn what they know about programming and data management on their own or the information is passed down within a lab, and as a result are unfamiliar with the equivalent of good lab practices for data science. The hidden costs this creates are significant: researchers spent weeks or months doing things that could be done in hours or days, do not know how trustworthy their results are, and are often unable to reproduce their own work, much less that of their colleagues.

[1] http://braembl.org.au/news/braembl-community-survey-report-2013
[2] http://elixir-uk.org
[3] http://www.katyjordan.com/MOOCproject.html

The workshop model to meet training needs

There are many challenges in providing effective training in data skills to researchers. One particular challenge is the substantial variation in the training occurring at institutions. There are many reasons for this. The curriculum is already full and there is not room to add specific courses or even lectures incorporating these topics. There are not instructors at a given institution who are able to teach these courses, either through a lack of knowledge or because of commitments to other activities. Additionally researchers are time-challenged. Existing commitments to research, grants and service often leave little time to develop new skills. Finally there is not currently a good model for community lesson development, so materials are often developed independently at each institution or department and there is not the opportunity for community engagement on what would be best taught or refinement as the lessons are taught multiple times. Ideally training would be high quality with materials vetted by practiced instruction, consistent across universities and locations, be able to be deployed at multiple and disparate locations, allow researchers to interact with the materials and the instructors and provide a relatively easy entry in to learning new topics.

A hands-on workshop model with community developed lessons is one that addresses these needs. A set of materials can be developed by the community that can share perspectives on best practices and taught broadly. This not only develops a more effective lesson, but because the same lessons are being taught multiple times there are opportunities for feedback and refinement of the lessons to deliver a higher quality product. The short, focused time of workshops gives researchers the committed time while attending the workshop to work on developing new skill sets. The hands on nature gives researchers the chance to develop their computational skills in the course of the workshop, so they leave with practical examples and hands on experience. Finally workshops can be taught by instructors from outside a given institution, so the institution does not have to rely on local knowledge or availability of instructors.

Software Carpentry[4] has been a leader in this approach and has been teaching best practices in software development to researchers with this format. It was created in 1998 as two-day intensive hands-on workshops to teach software practices fundamental to repeatability and accountability in scientific software development as well as strategies to be more effective and productive - version control, programming, software testing and the command line - enabling researchers to develop scripts or software that can reduce the time that things can be done from days and weeks to hours or days. All lessons are developed collaboratively with the community, and as with open source software, anyone can propose an improvement to its lessons (which are all freely available under a Creative Commons license). Those proposals are reviewed, improved, and finally merged into the core so that everyone can benefit from better explanations, examples, and exercises. All workshops are taught by volunteer instructors, more than 80 of which have been trained in the past year. Since 2010 alone, Software Carpenty has grown into a volunteer organization through which more than 160 instructors have taught two-day workshops for over 7000 people in 15 countries.

[4] http://software-carpentry.org/

Data Carpentry workshops to train researchers in data skills

Data Carpentry also uses this workshop model and was developed as a sister organization to Software Carpentry. Where Software Carpentry was developed to train researchers who were already programming, better software development practices, Data Carpentry is being developed to meet the needs of the everyday researcher who has data, big or small, that they need to analyze.

We have all experienced in our own work and in our interactions with colleagues how the inability to conduct an analysis or a lack of awareness of available methods limits research progress and leaves researchers feeling frustrated and disatisfied with their current data processing workflow. This sentiment was echoed by a survey of the researchers at the NSF BIO Centers, NSF funded centers focused on biological research, and was the original set of researchers we set out to target with training.

Given the effectiveness of the workshop model and the proven success of Software Carpentry, we developed Data Carpentry as two-day workshops to meet these data training needs and focus on standard steps in the data workflow - organizing, managing and analyzing data in a more efficient and reproducible way. Additionally, because people learn best when new skills are building on an existing framework, Data Carpentry workshops are designed to be domain specific so researchers can learn more quickly and effectively and see more immediately how to implement these skills and approaches in their own work. Also the workshops follow a narrative, using one domain relevant dataset throughout the workshop, and teaching the tools in the framework of addressing questions from that dataset.

We identified the following guidelines for the initial Data Carpentry core content:

- Workshops are domain specific. Each field has its own data types, analysis packages and standard problems to address. Being able to teach people in their domain allows them achieve two goals simultaneously: more immediately understand the questions and approaches, and 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 are 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. Social scientists, digital

humanists, biologists, librarians, and museum collections are all 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.

Data Carpentry progress and future plans

As our initial workshops focused on researchers in the NSF BIO Centers the first core workshop was developed with biological/ecological data - a survey of small mammals in a desert ecosystem. Using this dataset in the workshops we teach:

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

These workshops have now been taught seven times since May, 2014, with many more scheduled for 2015. We have an upocoming hackathon event to develop domain specific lessons for genomics and more organically integrate assessment into the workshops so that we can evaluate if learning objectives are being met. We have had broad interest in developing lessons in the social sciences, geosciences and neurosciences and are working with members of those communities to establish lessons in those domains. Additionally there has been interest in these workshops from librarians who are helping researcher manage their data lifecyles or are conducting their own analyses with digitized collections.

While our initial focus is on a core product for introductory workshops, we are planning to develop or incorporate more advanced topics such as Natural Lanaguage Processing, more advanced statistical topics, using cloud resources and using APIs for data access and sharing. In these topics and all workshop materials there will be a continued emphasis on conducting data and computation-heavy research more reproducibly and openly.

There are more hackathons planned to develop new and to improve the existing material. In collaboration with the aforementioned ELIXIR project Data Carpentry will co-run two of these events in the first half of 2015. ELIXIR is a large-scale European project focusing on providing "a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society."[5]. The hackathons will be immediately followed by a Data Carpentry workshop which will allow for testing the new materials and receiving feedback which will drive further work. Collaborating with such extensive projects like ELIXIR proves to be very beneficial for Data Carpentry. We are able to leverage the expertise, knowledge, infrastructure

[5] http://www.elixir-europe.org/

and network of our collaborator which enables us to promote and to scale up our initiative. In return, we provide ELIXIR with an established, tested and well received training model, together with a set of training materials and guidance on how to develop new ones. The ELIXIR Nodes[6] are also interested in growing their own pool of instructors and will provide means to run instructors training (train-the-trainer, as discussed in more details below).
Thanks to the efforts from the ELIXIR UK Node[7], the first Data Carpentry workshop in Europe was run in at the University in Manchester, UK in November 2014. The workshop received very positive feedback and we already had a number of inquiries about further planned events. Working together with ELIXIR we are now able to fulfill this growing demand. At the same time, in the long term, we are planning to empower ELIXIR and its collaborators making them more independent in running Data Carpentry at their local research organisations.
We are also working to establish relationships with foundations and industy. The goals of Data Carpentry to teach researchers skills for working more effectively with data is also aligned with many organizations. In particular, these skills help to enable the paradigm of data-driven discovery being advanced by the Gordon and Betty Moore Foundation, and the Moore Foundation has contributed to Data Carpentry to help with core organization and development.
More generally we are developing strategies to work with communities to develop content, respond to user input and solicit lesson contributions and recruit instructors to ensure a diverse, representative pool of contributors. Importantly Data Carpenty has also established a Code of Conduct for for both its workshops and participation in its community, and we are committed to providing safe, friendly environments to learn scientific computing.

Assessment of Data Carpentry effectiveness

The need for assessment in training activities is essential in order to ensure that learning objectives are being met and that the training is having a positive impact on the researcher's effectiveness and perceptions of their work. We have been conducting assemssment and are working to make it a fundamental component in to Data Carpentry training, both to determine short term outcomes (i.e. right after the workshop) and longer term outcomes (what is the impact of the training on the researchers work months or years after the training). Our approach to designing and using assesement is influenced by the experiences of Software Carpentry and the advice and approaches on assessment from those experiences in education assessment or in methodology of surveying different cohorts over time.
We are using two types of assessment: formal and informal. The formal assessment is conducted using well established methodologies such as questionnaire and (in the future) interviews. The formal assessment is thus well documented and allows for comparison between cohorts of learners and workshops. The questionnaires are fully digitalised which allows for quick and flexible manipulation of the collected data (it would be rather embarrassing if Data Carpentry struggled with analysing

[6] http://www.elixir-europe.org/about/elixir-nodes
[7] http://elixir-uk.org/

their own datasets!). The informal assessment is based on the discussions within the community. Whilst these discussions commonly happen via emails or shared online documents, the information is not methodologically collected and thus may typically serve as a supporting evidence, especially in any formal documents, such as grant proposals.
The assessment serves two purposes:

1. provides us with the feeback from the learners which then allows us improve Data Carpentry in many aspects, such as the mode of delivery of the workshops, the materials, the dissemination and outreach, and so on.
2. gives us evidence of the effectiveness and impact that Data Carpentry makes within various research areas and communities.

Whenever possible, we try to combine the assessment for both purposes. That is, we try to design our surveys so that their outcomes provide information and evidence serving both of the above goals. This means that the we do not overload the participants with too many surveys or other forms of assessment, but requires a deliberate approach to assessment planning described as described below.

Assessment to help improve Data Carpentry

For the formal assessment we use pre- and post-workshop questionnaires. The questionnaires are created using Qualtrics platform[8] which provides professional tools for running surveys. Qualtrics is used by a number of universities and research institutions in the United States. Thanks to the collaboration with iDigBio, we have been able to conduct our surveys via this platform. Thanks to help from Shari Ellis, an assessment expert at iDigBio, we were able to adapt the original Software Carpentry questionnaire for our own needs.
The pre-workshop questionnaire focuses on capturing the information about the relevant computational skills which the participants may already have before coming to the workshop. The questionnaire also includes a set of questions, or rather short specific tasks for manipulating data, and the participants need to rank their ability to complete these. We ask the participants about their expectations towards the training. This information allows us to learn more about our target audience which in turn helps us in planning on the necessary adjustments of the materials. The feedback included in the post-workshop questionnaire where the participants can freely say what they did and did not like as well as make suggestions for changes and improvements, provides this valuable perspective as well.
Since Data Carpentry is still in its early phase and the number of workshops run so far is not large, the informal feedback comes from direct interactions with the participants. The hands-on interactive nature of the workshops allows the instructors to closely observe the issues that the participants struggle most as well as note which bits of the training the students found particularily useful. The discussions with the students during the exercises and breaks are informative both

[8] http://www.qualtrics.com/

in terms of understanding their needs for training topics and their opinions about the lessons content.

Assessment for evidence of the impact made by Data Carpentry

The assessment which provides us evidence of the impact that Data Carpentry makes is possible mainly due to comparing the pre- and post-workshop questionnaires. Both questionnaire include a set of questions, or rather short specific tasks for manipulating data, and the participants need to rank their ability to complete these. By "relevant" we mean skills directly corresponding with the modules which we teach at the workshops. The participants are asked about their experience with programming languages and tools to manage and analyse data, as well as their everyday practices in working with data (such as data sharing, licensing and so on). The post-workshop questionnaire includes a similar set of questions as the pre-workshop one. This allows us to asses the direct impact that the workshops have by comparing what the participants said they could do before the training and after it.

A more challenging issue is to asses the long-term impact of the workshops. In oder to measure how the taught skills enable the participants to be more productive and effective in their research a more wide-scope study is needed. Software Carpentry has been trying to develop such assessment for at least a couple of years now. Three main difficulties are:

- problems in measuring and assessing that a given skills has actually contributed to the researcher's productivity and made their research more robust and reproducible;
- finding volunteers within the community who can fully commit to conducting such assessment, as in fact, it can be very time-consuming;
- pursuing the trainee cohorts to interview or survey them several weeks, months or even years after the training.

Therefore one of the Data Carpentry goals is to address the above challenges.

Building a community of instructors

Running an effective workshop means having instructors trained in how to teach, particularly in a workshop format. Software Carpentry has developed an effective train-the-trainers program based on pedagogy and experience[9]. Train-the-trainers programme is run either online for about six weeks (which involves mainly a lot of self-study work and regular bi-weekly conference calls) or during an intense two-day face-to-face event. The latter is a new mode of training the instructors and the plan is to hold it regularly at different locations enabling the training of more instructors.

[9] http://teaching.software-carpentry.org/

The curriculum for train-the-trainers has solid roots in research on best practices in education and aims to equip the future instructors with practical skills to effectively teach students who may come from different backgrounds and have different research goals. Data Carpentry instructors will also be trained in this way with additional focus on looking at Data Carpentry specific training modules.

Conclusion

Seven 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 enthusistic about hosting, teaching or taking these workshops as well as been engaged in the development 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 will not be able to teach researchers all the skills they need in two days, but we have shown that they are a way to get the process started and that we 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. Our goal is to empower researchers to be able to conduct the analyses necessary for their work in an effective and reproducible way.

Key components for continuing the progress of Data Carpentry and being able to offer workshops to researchers interested in developing these skills is to build a community of researchers who are developing, updating and improving lessons as well as a pool of instructors. We have found that many participants of Software and Data Carpentry workshops go on to be actively engaged in the community, particularly as workshop helpers and instructors. Running effective workshops means training these instructors in how to teach, particularly in a workshop format. Software and Data Carpentry have developed an effective train-the-trainers program based on pedagogy and experience. This train-the-trainers program along with the continued improvement and development of lessons, and the coordination of those activities, are important components of the scalability of the workshops and their continued ability to meet the ever-incresing demand for these skills. Additionally, there is the opportunity to deliver these workshops and lessons to researchers in emerging economies, allowing them to better participate in the analysis of publicly available data and to address questions of particular interest in their geographical regions and communities.

Workshops do scale more effectively than courses, but the in person aspect of workshops and requiring that there are instructors for each workshop does limit its scaling. The demand for Software and Data Carpentry workshops seems to be limitless: workshops fill within hours of being announced and more universities are requesting workshops than there are instructors available.

While our current focus is workshops, as data-driven approaches continue to become an even more fundamental components to research, we envision that data analysis courses and lessons will be integrated in to even the standard undergraduate curriculum, decreasing this specific workshop demand. We hope to be a part of that

transition as well and continue to be involved in the advancement of data literacy skills for researchers.