



Short-format Workshops Build Skills and Confidence for Researchers to Work with Data

Kari L. Jordan Ph.D., The Carpentries

Dr. Kari L. Jordan is the Director of Assessment and Community Equity for The Carpentries, a non-profit that develops and teaches core data science skills for researchers.

Marianne Corvellec, Institute for Globally Distributed Open Research and Education (IGDORE)

Marianne Corvellec has worked in industry as a data scientist and a software developer since 2013. Since then, she has also been involved with the Carpentries, pursuing interests in community outreach, education, and assessment. She holds a PhD in statistical physics from École Normale Supérieure de Lyon, France (2012).

Elizabeth D. Wickes, University of Illinois at Urbana-Champaign

Elizabeth Wickes is a Lecturer at the School of Information Sciences at the University of Illinois, where she teaches foundational programming and information technology courses. She was previously a Data Curation Specialist for the Research Data Service at the University Library of the University of Illinois, and the Curation Manager for Wolfram|Alpha. She currently co-organizes the Champaign-Urbana Python user group, has been a Carpentries instructor since 2015 and trainer since 2017, and elected member of The Carpentries' Executive Council for 2018.

Dr. Naupaka B. Zimmerman, University of San Francisco

Mr. Jonah M. Duckles, Software Carpentry

Jonah Duckles works to accelerate data-driven inquiry by catalyzing digital skills and building organizational capacity. As a part of the leadership team, he helped to grow Software and Data Carpentry into a financially sustainable non-profit with a robust organization membership in 10 countries. In his career he has helped to address challenging research problems in long-term technology strategy, GIS & remote sensing data analysis, modeling global agricultural production systems and global digital research skills development.

Tracy K. Teal, The Carpentries

Short-format workshops build skills and confidence for researchers to work with data.

Abstract

Training for data skills is more critical now than ever before. For many researchers in industry and academic environments, a lack of training in data management, munging, analysis and visualization could lead to a lack of funding to support sustainable projects. Today's researchers are often learning 'as they go' and need the flexibility of short, or self-paced learning experiences. Research results in educational pedagogy, however, stress the importance of guided instruction and learner-instructor interaction, which contrasts the need for 'just in time' training.

We've taken a distinctive approach to this problem, combining the power of guided instruction with the flexibility of short, focused learning experiences. Two-day, interactive, hands-on coding workshops train researchers to work with data, and have reached over 34,000 researchers, ranging from biologists to physicists to engineers and economists. Researchers have benefited from evidence-based teaching approaches to learning data organization (spreadsheets), cleaning (OpenRefine), management (SQL), analysis and visualization (R and Python).

This paper presents the long-term survey results showing the impact that short-format workshops have for increasing learner's skills and confidence in their coding abilities. Results show these two-day coding workshops increase researchers' daily programming usage, and sixty-five percent of respondents have gained confidence in working with data and open source tools as a result of completing the workshop. The long-term assessment data showed a decline in the percentage of respondents that 'have not been using these tools' (-11.1%), and an increase in the percentage of those who now use the tools on a daily basis (+14.5%).

Keywords: Assessment, data science, short courses

Introduction: State of Data Science Workforce Needs

Globally, data science talent is in high demand. In their widely cited report on big data, McKinsey Global Institute (MGI) estimated that, by 2018, in the United States, the shortfall in data science workforce would be 60% of its supply [1]. Although the term 'data science' was not in use in this 2011 report, the two occupational groups deemed key to creating value from data (i.e., "deep analytical talent" and "data-savvy managers and analysts") correspond to "data scientists" and "business translators," respectively, in a 2016 follow-up report [2].

The authors extend their overall findings to other high-income countries, while acknowledging that these display significant variations in the number, both gross and per capita, of their new graduates with deep analytical skills. In this way, a Canadian study [3] considered Canada's shortage in data talent to be the same as the United States' one determined by [1], proportionally to its population, while cross-matching it with a second estimate.

The more recent report, where the projection horizon is 2024, finds lower relative gaps between labour demand and supply [2]. Indeed, the projected supply and demand of data scientists in the

United States are 483,000 and 500,000 positions, respectively (only in a high-case scenario would demand reach 736,000 and, hence, yield a 50% shortfall). To add to the uncertainty, much of data science work is self-destructive: The automation of data preparation (supported by at least *some* machine learning) could lead to a shift in data science work, a decreasing demand for data science work, or both [2].

Regarding “business translator” roles, which combine data literacy with operational skills, more organizations are training existing workers, “building these capabilities from within”. In addition, 20% to 40% of new graduates in science, technology, engineering, and mathematics (STEM), business, and any field involving quantitative analysis would have to become these data-literate managers and analysts, in order to meet the United States demand of two to four million by 2024 [2]. The authors stress the importance of data visualization to support decision-making. To add to the complexity, some workers can and will take on more than one role, especially in small and medium-sized organizations.

What we have referred to as ‘workforce needs’ may be more correctly characterized as growth potential, in the sense that most industries are still capturing only a fraction of the potential value from data and analytics [2]. Beyond considerations about economic value and labour markets, data literacy is, in today’s and tomorrow’s data-driven world, a prerequisite for social participation [4]. Although our workshops target researchers, our approach is in line with the democratization of data literacy and education overall.

Data Science Training

With the booming commercial and academic market for data science and analytics training options, those seeking this training have a variety of choices. When we think of training offerings, we need to categorize what is being instructed and how this might fit into the process of a researcher adopting a new tool or skill into their current line-up. The training offered needs to fit the person’s needs for: awareness of tools, specific tool introduction, intensive practice, and skill mastery.

Short-form (1-2 hour) workshops are often the most universal offering for training. They are the easiest to book rooms for (or offer online as webinars), find instructors for, and create material for. For the participant, one hour is a reasonable amount of time to find in their day and there are rarely any follow-up requirements. Thus, there is very little risk of making a bad time investment for the learner, and the instructional team has a lot of flexibility in repeating the training and experimenting with content. From research methods to retirement plans, this format is an exceptional platform for learners to explore new tools and services. Even though hands-on practice can be quite limited in this format, this discovery and evaluation process is crucial.

Other short-form training options in a classroom-like setting include pre-conference workshops that run for 3-6 hours and week-long ‘summer school’ workshops. These will look very different depending on the audience, and have the time flexibility to include tool introductions with intensive hands-on practice, or provide more advanced training for commonly used tools.

These short formats are very common in the academic context and very suitable for institutions to provide locale-specific training, or for specific research communities to provide domain-specific training. Learners who have decided that a tool is worth further exploration and introduction will find these courses essential to move forward from a novice phase, or to reorganize some of their self-taught knowledge. Longer in-person trainings can be the ideal format from an instructional perspective. Learners can often get help installing the tools on their own machines, ask questions, and have the flexibility to work on their own projects during the course. The biggest limitation to these formats is because they are in person, and access is the most prohibitive factor. Access issues include scarcity of time to attend, ability to travel to and physically access the location of the event, and funding to cover the travel, lodging, and registration.

Online self-paced training often takes the form of these half to full-day trainings for individual modules. Depending on the site, these modules can be grouped into topical clusters, and then further into longer programs of study. As with any self-paced online program, being able to fit it into a busy schedule, complete it anywhere, and often for a much lower price provides strong benefits. However, completing everything online with static exercises means that learners do not always get the benefit of using the tools on their own computers or experience practicing with a full-scale realistic project.

Commercial multi-week boot-camps and academic programs of study have the strongest benefits of time, but have serious access issues. Most require physical access for several months or years, with hefty tuition costs. The length of time and focus on mastery is also inappropriate for learners wishing to become more skilled at a tool for research or a job. These longer formats are often for those who desire entry into a job field as a professional programmer, rather than an end-user programmer who will remain in their current position.

The reality is this: most trainings require that the learner spend time deciphering how to fit all these materials together to suit their needs and perform their own critical evaluation of the educational quality of these platforms and instruction sessions. The trainings often lack large-scale testing with multiple runs, open and community-driven modifications, transparent assessment, and ease of access to in-person training.

The Carpentries

The Carpentries (Software Carpentry (SWC) and Data Carpentry (DC)) is a non-profit organization that coordinates collective efforts to spread the skills necessary for data-driven research. DC curriculum delivers domain-specific hands-on intensive workshops covering the full lifecycle of data-driven research. Current workshops are designed for people with little to no prior computational experience. These workshops include data organization and cleaning, data management and scripting for data analysis and visualization to empower researchers to work more effectively and reproducibly with data [5]. SWC curriculum focuses on common approaches for automating repetitive tasks, data management, defensive programming, source control, and code execution environments [6]. These workshops use active learning to give learners the opportunity to try the skills and gain immediate feedback with low learner-teacher and helper ratios, guided and independent exercises, and an interactive and friendly environment

for learning [7]. Combined, these workshops provide a strong opportunity to learn and the needed conceptual framework for participants to excel in data-driven research [8] - [9]. The Carpentries have built a network of instructors who are themselves researchers and have expressed significant career benefits from volunteering to teach workshops [10].

Tools we Teach

The Carpentries teach workshops around a base programming language, either Python or R. Python and R are both general-purpose, high-level, interpreted, multi-paradigm, cross-platform, free/libre and open-source (FLOSS) programming languages. These general features make these languages installable by any learner, accessible to novices, and transferrable to other applications. Specifically, relevant to scientific computing and data analysis are the rich ecosystems which have developed upon both these languages. These ecosystems consist of many software packages, standard practices and, last but not least, active communities. Thus, we teach tools which are not only effective at solving problems found in scientific and data-driven research, they are also popular, growing, and conducive to sharing, collaboration, and continuous improvement. We could make a similar case for SQL (Structured Query Language), which we teach in both a DC lesson and a SWC lesson. This tool is a standard and key building block for databases. DC workshops start off with a lesson on data organization, where we use spreadsheet software (such as LibreOffice or Microsoft Excel). Such software is usually familiar to learners who have no coding experience. Then, we teach data cleaning with OpenRefine, a FLOSS program which lets users explore, correct, and format data (refining raw data). On a meta level, this software represents a smooth transition from working with graphical interfaces to writing commands and scripting. SWC workshops teach automation with the Unix shell, a very traditional and fundamental tool. It lets the user interact with the computer at a base level, whether by running commands interactively or by scripting. In SWC workshops, we teach version control of source code with Git, a *de facto* standard. We also teach GitHub, a popular Git-based web service for hosting code repositories and facilitating contribution and collaboration. Meanwhile, we warn learners that their research organizations might recommend or require them to use other hosting platforms.

The Carpentries Teaching Practices

Carpentries instructors are trained to deliver content using a variety of evidence-based teaching practices. These techniques include sticky notes, learners using their own machines, minute cards, one up/one down, collaborative note-taking (Etherpad), live coding, formative assessment, Code of Conduct, and pre- and post-workshop surveys. A summary of each is provided below.

Sticky notes

Colorful sticky notes are used as status flags throughout workshops. Instructors and helpers float around the room identifying who needs assistance as learners place a sticky note (e.g., red) on their computer. In this way, learners do not need to raise their hand and draw attention to themselves, and the lead instructor is able to continue delivering the material.

Learners use their own machines

It is important for learners to leave the workshop with their own machine set up to do real work. This is why we ask learners to bring their own machines to the workshop. Our instructors teach on three major platforms: GNU/Linux, Mac OS X, and Windows.

Minute cards

Twice per day during a workshop, *minute cards* are collected to get anonymous feedback from learners. Learners are asked to write one positive aspect about the content and deliver of the workshop content, and one negative (or area of improvement). Instructors are encouraged to change the feedback prompt to elicit different types of feedback at each break. Instructors review the minute cards to look for patterns, and address commonly raised issues with the workshop participants. In this way, misconceptions are addressed, and instructors are able to adjust their delivery as needed to ensure learners understand the material.

One up/One down

Learners are asked to give feedback at the end of each day using a technique called “one up, one down”. Alternately, learners give one positive and one negative point about the day, without repeating a point that has already been discussed. In this way, learners are encouraged to speak about what they truly think once all the “safe” feedback has been shared. Instructors write down their feedback without commenting while collecting it. Instructors then discuss their feedback and how they explicitly plan to act on it.

Collaborative note-taking

The use of collaborative note-taking on either an Etherpad or Google document offers an avenue for dialogue and questions to be raised throughout the workshop. Learners are able to share links and bits of code, and the document remains available after the workshop.

Live coding

Rather than displaying a slide deck of code, instructors deliver content by live coding. This method provides learners the opportunity to practice, and receive continuous feedback about their code. It is important to keep in mind, however, that feedback is not helpful if you cannot understand it. Live coding facilitates tacit knowledge (i.e., learning by watching *how* instructors do things).

Formative assessment

Formative assessment, or diagnostic testing, allows instructors to modify teaching and learning activities to improve student learning [11]. The Carpentries’ lesson materials include challenge questions (exercises) to help learners evaluate their level of knowledge of the tools covered in the workshop. These challenge questions also offer opportunities for instructors to address misconceptions in content knowledge.

Code of Conduct

The Carpentries are committed to making participation in our workshops a harassment-free experience for everyone, regardless of level of experience, gender, gender identity and expression, sexual orientation, disability, personal appearance, body size, race, ethnicity, age, or religion [6]. We establish norms for interaction by having, discussing, and enforcing a Code of Conduct such that our workshops provide open and inclusive learning environments.

Pre- and post-workshop surveys

Learners complete surveys before and after attending a workshop. These surveys include questions to help instructors understand their attendees' prior experience and backgrounds before the workshop starts. Using this information, instructors can decide how they will approach the material and which exercises are likely to be appropriate for their learners. Additionally, the surveys measure confidence in using the tools covered, and self-efficacy.

Workshop Impact

An analysis of DC's workshop impact [12] showed workshops have significant impact on learners. Learners had high means for research computing efficacy, and expressed satisfaction with their instructional team, as mentioned in this quote from a learner:

“It introduced basic knowledge of data management using robust software and platform. The instructors were well prepared, knowledgeable, very helpful and created an interactive environment to make learning of the skills easier. Overall, they were very enthusiastic in what they were doing.”

Similar results from the analysis of SWC's workshop impact [13] strongly suggested that workshops provide “a welcoming environment for its learners” where the material not only matches the workshop description, but is worth the time learners spend learning it. These respondents identified acquiring skills they are able to apply immediately in their research and/or job function.

As post-workshop survey results have been positive, we are interested in what long-term effect these workshops are having on learners six months or more after a workshop, and what lessons learned we can offer ASEE as growth opportunities for its Computing & Information Technology Division.

Method

Our goal was to measure the long-term effects our workshops have on learners six months or more after attending a workshop. We are interested in learners' confidence and motivation to use the tools they learned, and specific behaviors learners adopted after attending a Carpentries workshop.

We compiled existing instruments measuring computer self-efficacy [14], Java programming self-efficacy [15], Python and computational ability [16], self-efficacy towards FLOSS projects [17], and student-instructor relationships [18]. Assessment specialists on staff and from our instructor community used a rubric to vote on whether to omit questions, keep them as-is, or adapt them for the purposes of our data collection. Rather than focusing on learners' skills with respect to particular tools, we wanted to focus on assessing learner confidence, motivation, and adoption of good research practices [19], as these elements represent the primary goals of our workshops. Confidence and motivation promote community building, a significant focus area of The Carpentries.

The final survey instrument included 26 questions. Figure 1 provides a select few questions from the survey. The entire survey, data set, and code used to prepare this paper can be found on our GitHub repository at <https://github.com/kariljordan/ASEE>.

The statements below reflect ways in which completing a Carpentry workshop may have impacted you. Please indicate your level of agreement with the statements

- I have been motivated to seek more knowledge about the tools I learned at the workshop.
- I have made my analyses more reproducible as a result of completing the workshop.
- I have improved my coding practices as a result of completing the workshop.
- I have gained confidence in working with data as a result of completing the workshop.

Figure 1: Select questions from long-term assessment survey.

Survey collection

The survey was distributed twice: in March 2017 and October 2017. The survey was distributed via electronic mail (MailChimp) using SurveyMonkey to learners in our database who opted in to receive communications from us. Though our database includes nearly 9200 email records, many of the emails were flagged as spam. Eighty percent of the emails sent were delivered before we were flagged for a high bounce rate. Many of our learners use academic emails that go away when they leave their institutions. We recorded 504 total responses (476 in March, 28 in October). We did not send follow-up emails for either the March or October cohort.

Data cleaning

A CSV file of the data collected was downloaded from SurveyMonkey and uploaded to OpenRefine for cleaning. Unique identifiers (i.e., IP address, workshop location, and open-ended responses) were removed as our data is open and available on online repositories. These stipulations are in place as part of our Institutional Review Board (IRB) agreement.

Limitations

The primary limitation for this instrument and analysis was the assessment purpose. This meant that the instrument would naturally be shorter, the questions not designed for hypothesis testing, lack of original piloting, and incomplete metadata about survey distribution. The questions were adapted from other inventories that were not all intended to be broken up. Given that our purpose was not for instrument replication or experimental group testing, we do not anticipate this limitation having a significantly negative impact on their use to assess the impact and usefulness of our curriculum. As this was community developed, there may be unanticipated design limitations to using mixed instruments.

As mentioned in previous sections, the true response rate can only be estimated because the total invitation count has been lost from the MailChimp interface, as the information is only available for thirty days after sending a mass email. The two cohorts are unbalanced, with the March cohort spanning individuals who took a workshop on or before December 1, 2016, and the October cohort consisting of workshop participants from the previous six months. The October cohort invitation size is estimated to be 10% of the March cohort size. This may impact some responses, as a respondent may not have a complete memory of their practices and changes made directly after the workshop.

We also know that not every workshop participant received the survey invitation, and the nature of open access instructional material means that there will always be alternate format uses that the organization will never be aware of. The biggest group that we know of are institutions that run self-organized workshops. Some of these are branded with the respective Carpentry name, and thus we are usually able to make contact with the instructors to get their information. However, participant information is not always sent to program coordinators. There are additional areas where a learner's information may be missed. Some instructors may not leave time or request learners to complete the pre-workshop survey, which is how we gather this contact information. Learners may choose to not complete the post-workshop survey, or opt out of receiving future communication.

The survey completion rate was 81%. This is because demographic questions (race/gender) are only asked to respondents who took a workshop in the United States. Additionally, we had optional open-ended response questions at the end of the survey that many individuals did not answer.

Our pool of learners are distributed internationally, meaning that academic status groups may not match our selections. Our primary audience is within higher education, but there may be styles of academic employment or student status that cannot be cross-walked easily into the labels we offered. Only 28 respondents selected the "Other Academic Staff" option for this question. This is also of particular concern with the question asking for a research domain. These large groupings are divided differently for some educational traditions and countries. Our survey was also only available in English, which may have influenced both the interpretation of some question text, but also may have caused an unknown number of interested respondents to not attempt the survey.

Our learners have a wide variety of reasons for taking a workshop. Many equally motivated learners will have unequal ability to make changes to the tools, programming languages, and practices within their research. This means that not all participants have equal reason or capability for changing or adopting something new into their workflow. This capacity to make a change was not measured, so we do not know the impact of this effect.

This instrument also changed from the March cohort to the October one. No questions were removed or had the text changed, but four questions about first-time survey completion, time since attending a previous Carpentries workshop, gender identity, and racial/ethnic identity were added. This means that number of responses for those new questions will have a much lower response rate. Overall, the instrument is already quite short with 28 questions in the March version and 32 in October. This is too short to establish full effect power for many of the source instruments, but this was not the purpose of the design. While this length was ideal for the context and population of respondents, it does mean limited scope for analysis and baseline comparisons of the source instruments.

Analysis and Results

The analysis in this paper was completed in RStudio collaboratively using several R packages including knitr, tidyverse, likert, and testthat:

- knitr was designed for dynamic report generation in R;
- tidyverse is a collection of R packages for data science;
- likert is an R package used for analyzing and visualizing Likert items;
- testthat is a popular unit testing package.

Respondent Demographics

The majority of survey respondents completed a workshop in the United States (48.5%), with 12.2% completing a workshop in Canada, and 7.5% completing a workshop in the UK. Other responses included Australia (4.8%), Norway (3.4%), South Africa (2.7%) and New Zealand (1.7%). Figure 2 provides a breakdown of survey respondents by country.

As our workshops are targeted toward researchers, the majority of respondents (35%) were graduate students (Figure 3), with 19% identifying as postdocs, 13% academic research staff, and 12% faculty and industry.

Table 1 provides a breakdown of respondents' field of research, work, or study. The majority (36.2%) of respondents work in life sciences, which makes sense as our DC workshops are domain specific for Ecology and Genomics. There is also representation in engineering (8.7%), biomedical/health sciences (21.5%), and mathematics/statistics (8.9%).

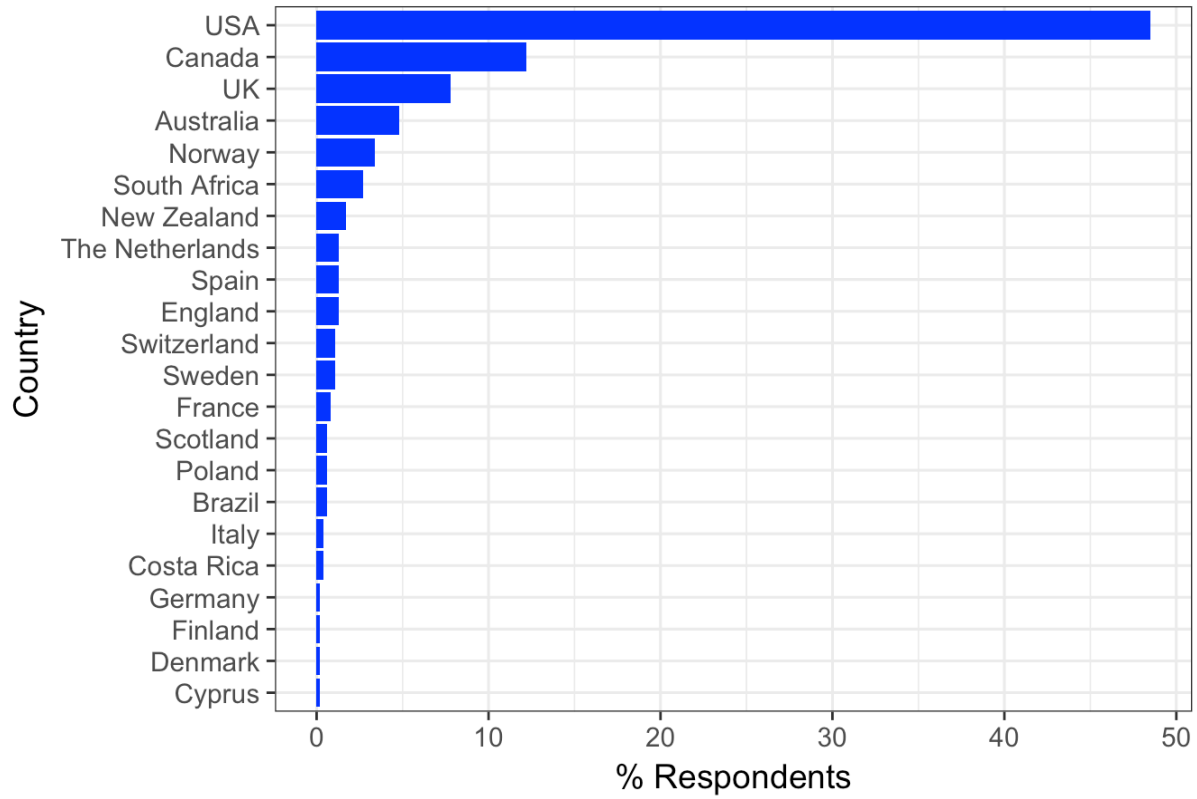


Figure 2: Breakdown of Respondents by Country

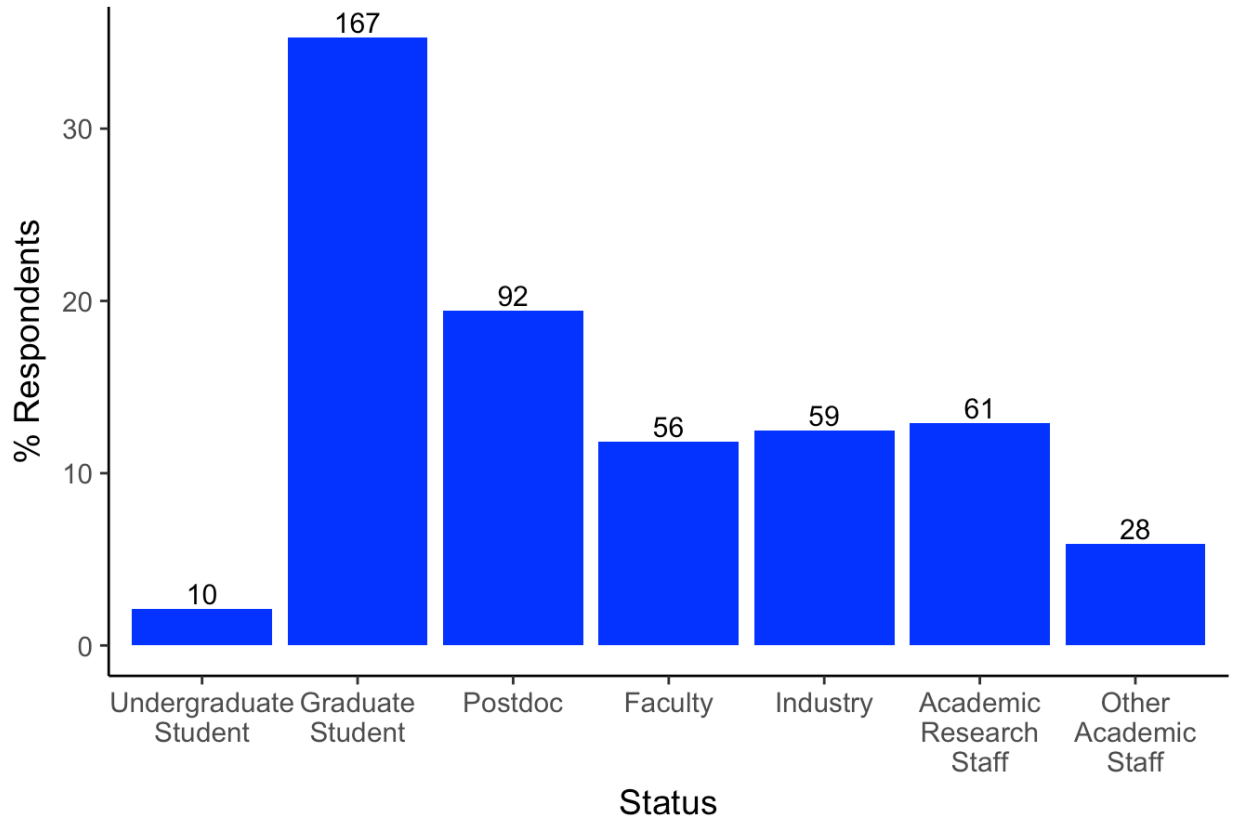


Figure 3: Respondents' Status

Table 1: Respondents' field of research, work, or study*

| Field | n | % |
|--|----------|----------|
| Life Sciences | 175 | 36.2 |
| Biomedical/Health Sciences | 104 | 21.5 |
| Agricultural or Environmental Sciences | 81 | 16.8 |
| Physical Sciences | 57 | 11.8 |
| Earth Sciences | 50 | 10.4 |
| Mathematics or Statistics | 43 | 8.9 |
| Engineering | 42 | 8.7 |
| Computer Science | 38 | 7.9 |
| Social Sciences | 25 | 5.2 |
| Library Sciences | 21 | 4.3 |
| Humanities | 14 | 2.9 |
| Business | 6 | 1.2 |

**Respondents were asked to check all that apply.*

Tools Covered

As mentioned earlier, the tools we teach in our workshops include base data cleaning and organization, data analysis and visualization with base programming languages, and version control. Table 2 provides the most frequent combination of tools covered in workshops attended by respondents to the survey.

Table 2: Most frequent combination of tools covered

| Tool | n | % |
|------------------------------|----------|----------|
| Git, Python, Unix Shell | 96 | 19.0 |
| Git, Python | 41 | 8.1 |
| Git, Python, SQL, Unix Shell | 39 | 7.7 |
| Git, R, Unix Shell | 31 | 6.2 |
| Git, R | 22 | 4.4 |

| | | |
|----------------------------|----|-----|
| Git, Python, R, Unix Shell | 19 | 3.8 |
| Git, Python, SQL | 16 | 3.2 |
| Git, R, SQL, Unix Shell | 12 | 2.4 |

Programming usage pre/post-workshop

Respondents were asked how often they use programming languages (R, Python, etc.), databases (Access, SQL, etc.), version control software and/or the shell before completing a Carpentries workshop, and since completing a Carpentries workshop. Understanding respondents’ programming usage both before and after attending a Carpentries workshop was one goal of this survey. Our hope is that the workshops favorably influence use of programming tools learned. Figure 4 shows a comparison of how often respondents use the tools they learned before attending the workshop, and more than six months after completing a workshop. We see an increase in respondents using the tools they learned on a daily basis as a result of completing a workshop. Additionally, the chi-squared residuals of post-workshop frequencies show that respondents are programming more often (Figure 5).

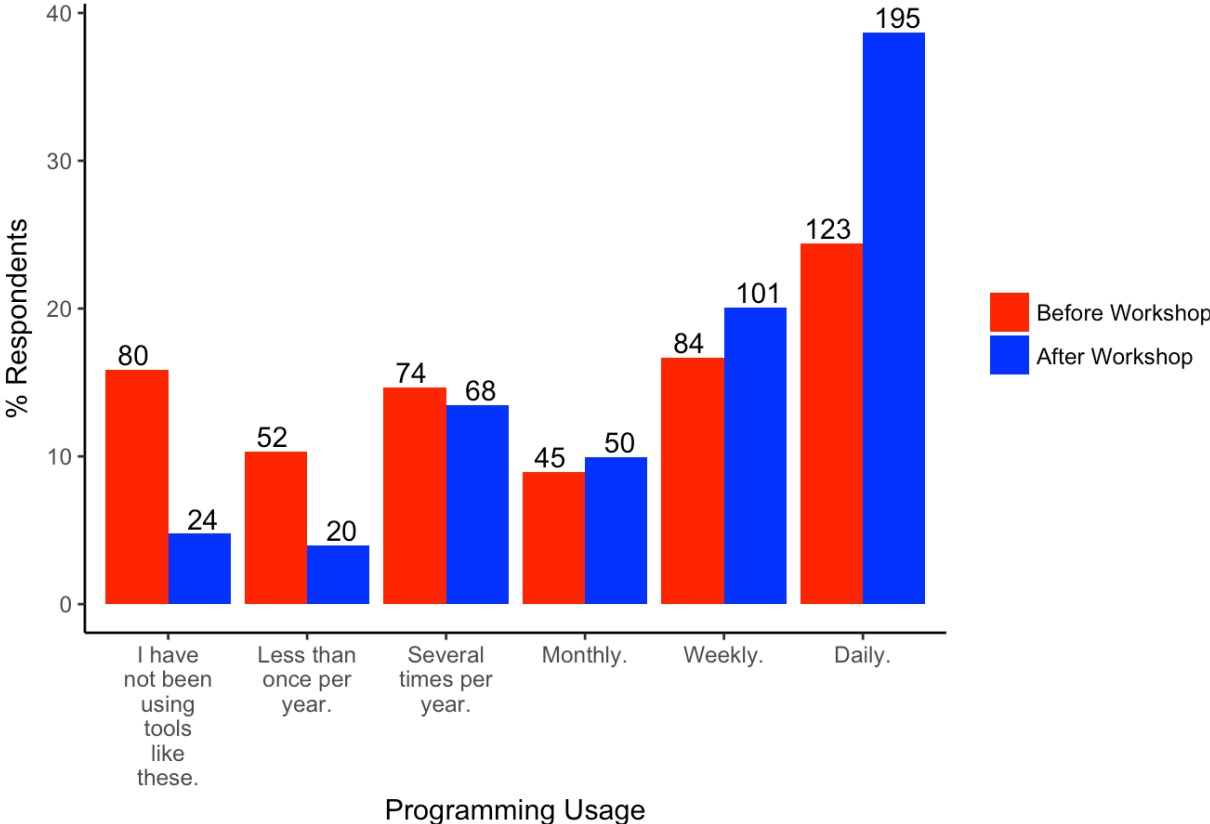


Figure 4: Respondents’ Programming Usage Increased

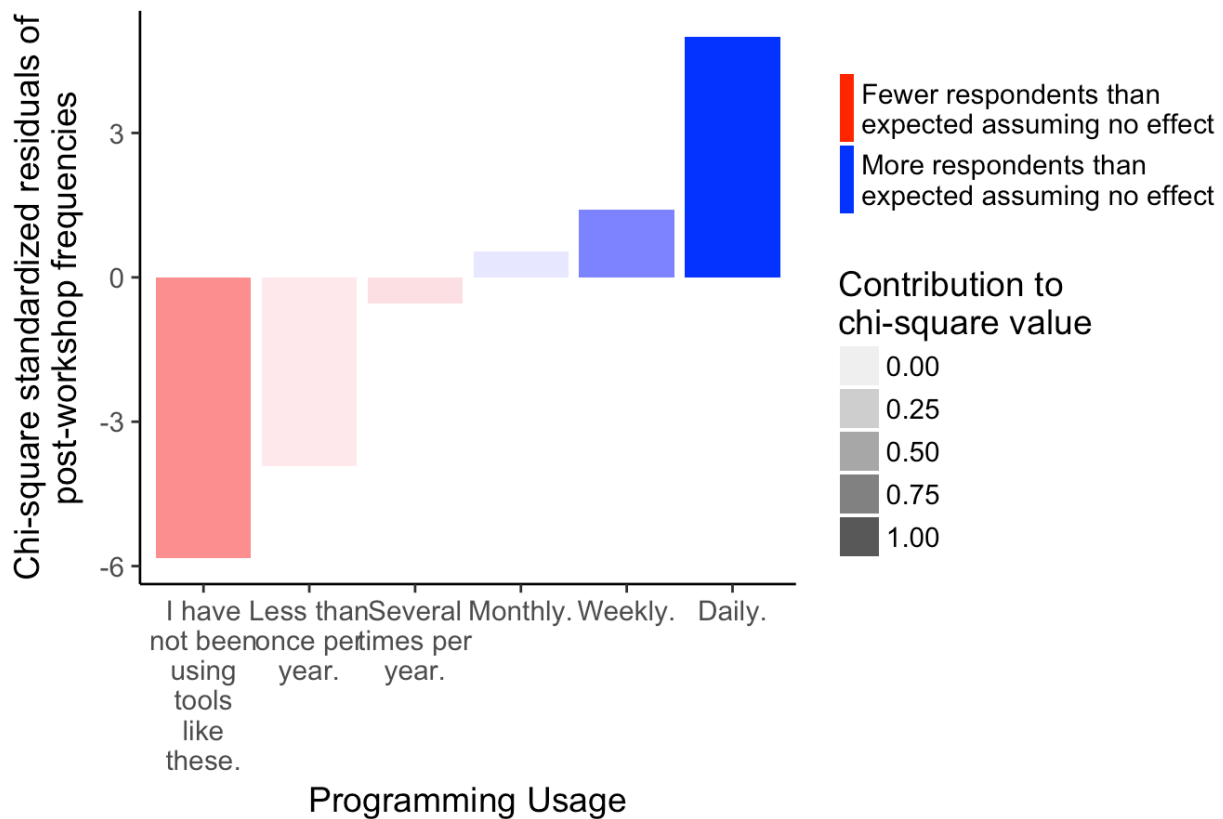


Figure 5: Respondents program significantly more often

Workshop Impact

Figure 6 shows respondents' perception of workshop impact on several factors, including career, confidence, and continuous learning. Respondents were asked to rate their level of agreement (1-Strongly disagree to 5-Strongly agree) with the statements below. The x-axis labels for the figure are in bold, and correspond to the statement following.

- **Reproducible:** I have made my analyses more reproducible as a result of completing the workshop.
- **Recognition:** I have received professional recognition for my work as a result of using the tools I learned at the workshop.
- **Productivity:** My research productivity has improved as a result of completing the workshop.
- **Motivation:** I have been motivated to seek more knowledge about the tools I learned at the workshop.
- **Confidence:** I have gained confidence in working with data as a result of completing the workshop.
- **Coding:** I have improved my coding practices as a result of completing the workshop.
- **Career:** I have used skills I learned at the workshop to advance my career.

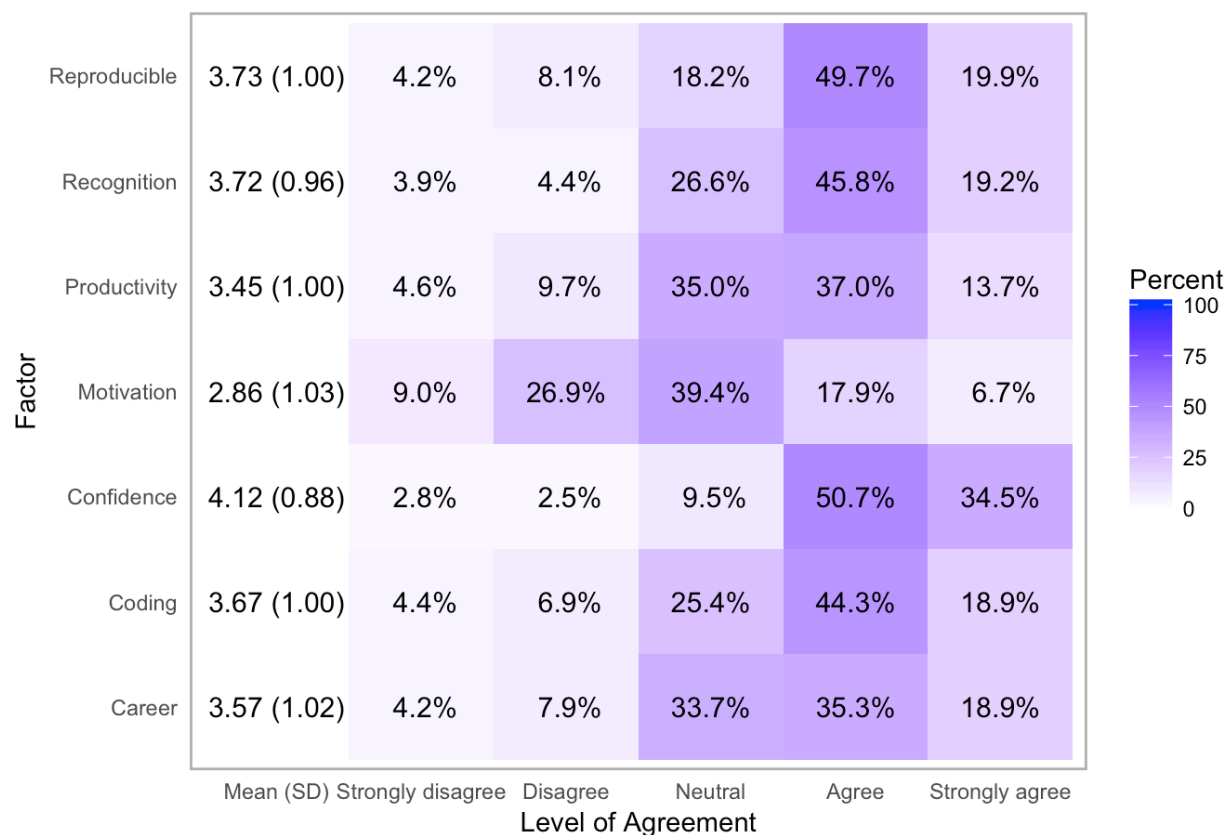


Figure 6: Perception of Workshop Impact

Behaviors Adopted

Respondents were asked to identify the behaviors they adopted as a result of completing a Carpentries workshop. Sixty-percent of respondents use base programming languages and/or the command line to automate tasks, and more than forty-percent have improved their data management and project organization, reuse code, and use version control to manage code. Additionally, the majority of respondents are **more confident** now in using the tools than before they completed the workshop (Figure 7).

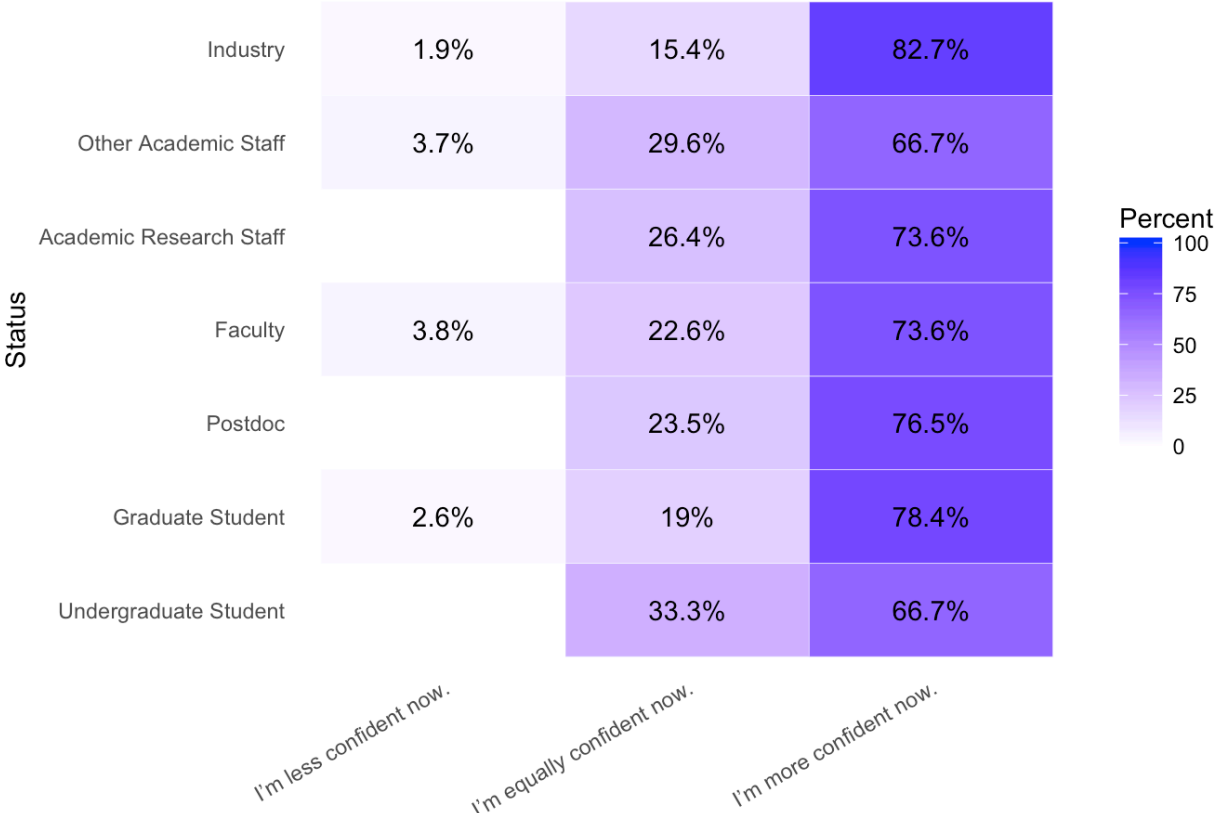
Table 3: Behaviors Adopted Post-Workshop*

| Behaviors Adopted | n | % |
|--|-----|------|
| Using programming languages like R or Python, or the command line to automate repetitive tasks | 274 | 66.2 |
| Improving data management and project organization. | 204 | 49.3 |
| Using version control to manage code. | 191 | 46.1 |

| | | |
|--|-----|------|
| Reusing code. | 180 | 43.5 |
| Sharing code or data publicly on places like GitHub or FigShare. | 128 | 30.9 |
| Using databases, scripts and queries to manage large data sets. | 124 | 30.0 |
| Using version control to collaborate online (in public or private repositories). | 123 | 29.7 |
| Transforming step-by-step workflows into scripts or functions. | 118 | 28.5 |
| Developing a data management and analysis plan. | 75 | 18.1 |

*Respondents were asked to check all that apply.

Change in confidence



Change in Confidence by Status

Figure 7: Change in Confidence by Status

Usage of tools for research/work

We have identified specific outcomes directly related to research and/or work, and asked learners if they had achieved these outcomes six months after completing a workshop. Respondents (more than 50%) reported that the tools they learned improved their overall efficiency, as well as their ability to manage and analyze data (Table 4).

Table 4: How have tools learned helped in research/work?*

| How Tools Have Helped Research/Work | n | % |
|--|----------|----------|
| They are improving my overall efficiency. | 254 | 59.5 |
| They are improving my ability to analyze data. | 228 | 53.4 |
| They are improving my ability to manage data. | 214 | 50.1 |
| I am not using the tools I learned. | 65 | 15.2 |
| The tools I learned have not helped me with my work. | 30 | 7.0 |

**Respondents were asked to check all that apply.*

Continuous Learning

One key objective of The Carpentries is that our learners continue to improve their skills after completing a workshop, whether by joining a local community, or using self-guided materials. This can take many forms, including participating in short courses (in-person and online). We asked respondents to tell us which learning activities they have participated in since completing a workshop. The majority of respondents (35%) have used non-Carpentries, self-guided material, and 19% responded having used Carpentries' self-guided material.

Table 5: Learning Activities*

| Learning Activities | n | % |
|---|----------|----------|
| Used non-Carpentry self-guided material. | 135 | 35 |
| Used self-guided Carpentry lesson material. | 72 | 19 |
| Participated in an in-person short course. | 63 | 17 |
| Participated in an online short course. | 48 | 13 |
| Participated in a Meetup. | 35 | 9 |
| Participated in a semester long course. | 28 | 7 |

**Respondents were asked to check all that apply.*

Summary and Growth Opportunities

Though we've identified several limitations with respect to data collection and sample size, we see from the results of this analysis that respondents are adopting positive behaviors six months or more after attending a Carpentries' workshop. We believe these results stem from our commitment to creating open and collaborative curriculum and teaching environments. As all of

our lessons are available online, we offer learners ease-of-access to training material, and promote sustainable learning.

We believe our teaching approaches are the reason why our results are so positive. Interactive teaching approaches are becoming an established norm, especially for computing and other hard sciences [20]. By promoting a supportive environment, embracing mistakes, and building community, we are able to make learning programming less intimidating.

We recognized that thirty-five percent of respondents go on to use non-Carpentry self-guided material. We believe this is because many of our lessons are domain specific, and our material serves as a starting point. Non-Carpentry self-guided materials give learners the opportunity to delve into more detail once they've learned the basic computing skills we teach. We introduce fundamental concepts, and help learners enrich their vocabulary such that they can seek material relevant to specific tasks they want to accomplish.

As the Computing & Information Technology division is “concerned with meeting the information crisis, both through efficient use of available information tools and through development of better systems,” we offer this work as a sample model for developing educational resources for teaching and learning in computing.

In no way have we achieved perfection in delivering effective short-format data science training. We are always growing as instructors and as curriculum developers. Workshop participants usually come with a challenging mix of backgrounds and skills. In the classroom, instructors try to accommodate the diversity of needs by leveraging our interactive teaching methods (sticky notes, minute cards, Etherpad, formative assessment, etc.). This nearly real-time adjustment is no silver bullet though. Common pitfalls include: failing to demonstrate the innocuity of typing something ‘wrong’; engaging in “what if” conversations at the expense of the hands-on practice; getting side-tracked by (valid, interesting) questions from one participant (or a small group). We strive for continual improvement in the longer run as well. To this end, we encourage instructors to pay attention to pre- and post-workshop surveys (especially free-form comments) and to participate in instructor discussion sessions. Directly or indirectly, these typically drive additions and changes (arguably, improvements) to our lesson materials. As an invitation to the broader community of data science educators, we welcome discussions and ideas on training challenges.

Acknowledgement

This work was funded in part through a grant from the Gordon and Betty Moore Foundation.

References

[1] McKinsey Global Institute, *Big data: The next frontier for innovation, competition, and productivity*, June 2011. Accessed at https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf on Wed, December 27, 2017.

[2] McKinsey Global Institute, *The age of analytics: Competing in a data-driven world*, December 2016. Accessed at

<https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.ashx> on Wed, December 27, 2017.

[3] Canada's Big Data Consortium, *Closing Canada's Big Data Talent Gap*, October 2015. Accessed at <https://ucarecdn.com/a342af00-e519-4dbb-a81b-3e93267b345b/> on Wed, December 27, 2017.

[4] M. B. Twidale, C. Blake, and J. Gant, "Towards a data literate citizenry," paper presented at the *iConference 2013, Fort Worth, TX. Feb. 2013*. Accessed at <https://www.ideals.illinois.edu/bitstream/handle/2142/38385/189.pdf> on Thu, December 28, 2017.

[5] Data Carpentry. January 2013. Retrieved from <http://www.datacarpentry.org/lessons/>

[6] Software Carpentry. January 2013. Retrieved from <https://software-carpentry.org/conduct/>

[7] T. K. Teal, K. A. Cranston, H. Lapp, E. White, G. Wilson, K. Ram, and A. Pawlik, "Data Carpentry: Workshops to Increase Data Literacy for Researchers," *International Journal of Digital Curation* 10(1):135-143, 2015.

[8] G. K. Sandve, A. Nekrutenko, J. Taylor, E. and Hovig. "Ten simple rules for reproducible computational research," *PLoS Comput Biol*, 9(10):e1003285, 2013. Adapted from <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285>

[9] J. M. Osborne, M. O. Bernabeu, M. Bruna, B. Calderhead, J. Cooper, N. Dalchau, S-J. Dunn, A. G. Fletcher, R. Freeman, D. Groen, B. Knapp, G. J. McInerny, G. R. Mirams, J. Pitt-Francis, B. Sengupta, D. W. Wright, C. A. Yates, D. J. Gavaghan, S. Emmott, and C. Deane, "Ten Simple Rules for Effective Computational Research," *PLoS Computational Biology*, 10(3):e1003506, 2014.

[10] B. M. Duckles, "Value of Software Carpentry to Instructors Report," Report to Software Carpentry Foundation and Data Carpentry Foundation, 2016.

[11] J. Dolin, P. Black, W. Harlen, and A. Tiberghien, "Exploring Relations Between Formative and Summative Assessment," in *Transforming Assessment. Contributions from Science Education Research*, J. Dolin and R. Evans (eds), vol. 4, Springer, Cham, 2018.

[12] K. L. Jordan, "Data Carpentry Assessment Report: Analysis of Post-Workshop Survey Results." October 2016. <http://doi.org/10.5281/zenodo.165858>

[13] K. L. Jordan, B. Marwick, J. Duckles, N. Zimmerman, and E. Becker, "Analysis of Software Carpentry's Post-Workshop Surveys," July 2017. <http://doi.org/10.5281/zenodo.1043533>

- [14] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS quarterly*, Vol. 19, No. 2, pp. 189-211, Jun. 1995.
- [15] P. Askar and D. Davenport, "An investigation of factors related to self-efficacy for Java programming among engineering students," *TOJET: The Turkish Online Journal of Educational Technology*, Vol. 8 No 1, 2009.
- [16] C. T. Brown, "Python and Computational Ability Scale," STEM E.D., LLC, Davis, CA, 2012.
- [17] I. Steinmacher, I. Wiese, T. U. Conte, and M. A. Gerosa, "Increasing the self-efficacy of newcomers to Open Source Software projects," in *2015 29th Brazilian Symposium on Software Engineering (SBES)*, pp. 160-169, IEEE, September 2015.
- [18] G. Creasey, P. Jarvis, and E. Knapcik, "A measure to assess student-instructor relationships," *Center for Teaching, Learning, & Scholarship at Georgia Southern University, Statesboro, Georgia, USA*, 2009.
- [19] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, "Good enough practices in scientific computing," *PLoS Comput Biol*, 13(6): e1005510, 2017.
<https://doi.org/10.1371/journal.pcbi.1005510>.
- [20] Lungeanu, Diana, Rochelle E. Tractenberg, and George I. Mihalas. "Developing an interactive approach in teaching medical informatics." *Studies in health technology and informatics* 124 (2006): 885-890.