Data Carpentry: Teaching Reproducible Data Driven Discovery

Tracy K. Teal, PhD  
@tracykteal, @datacarpentry

François Michonneau, PhD  
@fmic_

Mine Çetinkaya-Rundel, PhD  
@minebocek
Our increasing capacity to collect data is changing how we can look at the world.
Software and tools allow us to turn data into information.
People turn information into knowledge.
We can unleash the potential of data by empowering people.
It’s challenging to manage, store, analyze, or even view this data.
Researchers view the major limiting factor in research progress as a lack of expertise in how to handle and analyze data.

How do we scale data skills along with data production?
Build local communities and capacity
Software and Data Carpentry

Building community teaching universal data literacy
Grassroots training effort

- Developed by practitioners for practitioners

- Identify skills and best practices needed in software development, data management, and analysis in given domains

- Collaboratively and iteratively developed and openly licensed (CC-BY) training materials (all available on GitHub)

- Organize and deliver two-day, intensive hands-on workshops in fundamental data analysis skills using a pool of volunteer helpers and instructors
Software and Data Carpentry

- Core skills for effective research computing
- Two-day hands-on workshops
- Collaboratively developed, openly licensed lesson materials
- Over 800 trained volunteer instructors on 6 continents
Software & Data Carpentry workshops

We know we can’t teach everything in two days, but the goal is to teach foundational skills to reduce the activation energy for getting started and for people to know what’s possible.
Workshop goals

- Primary goal is increased confidence in a learners ability to do computational work and continue to learn more.
- Lower the activation barrier for people to get started working with data. Often people don’t know where to start.
- The skills we teach should be ones they can immediately apply to their research.
- People should learn the things we’re teaching.
- We should see a shift in perspective in the value of skills like scripting for better research and reproducible research.
- People should have a positive workshop experience that empowers them to do better work with data.
Content
- Real database examples
- Access to new tools
- Very detailed introduction

Delivery
- Easily to follow and interactive
- Hands-on exercises
- Straightforward for beginners
- Ability to ask questions without slowing pace

Instructors
- Patient instructors
- Qualified facilitators
Hands-on intensive workshops

- Two days
- Hands-on
- Qualified instructors
- Helpers
- Post-it notes!
- Friendly learning environment
Collaboratively developed curriculum

<table>
<thead>
<tr>
<th>layout</th>
<th>title</th>
<th>subtitle</th>
<th>minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>page</td>
<td>Programming with R</td>
<td>Analyzing patient data</td>
<td>30</td>
</tr>
</tbody>
</table>

**Learning Objectives {.objectives}**

- Read tabular data from a file into a program.
- Assign values to variables.
- Select individual values and subsections from data.
- Perform operations on a data frame of data.
- Display simple graphs.

We are studying inflammation in patients who have been given a new treatment for arthritis, and need to analyze the first dozen data sets. The data sets are stored in **comma-separated values (CSV)** format. Each row holds the observations for
- Focused on data - teaches how to manage and analyze data in an effective and reproducible way.
- Initial focus is on workshops for novices - there are no prerequisites, and no prior knowledge computational experience is assumed.
- Domain specific by design – currently have lessons in ecology, in genomics developed with iPlant, in geospatial data developed with NEON, and on APIs with rOpenSci.

**Curriculum**
- Project and data organization
- Data cleaning and quality control
- Data analysis and visualization in R or Python

**Domain dependent**
- Introduction to cloud computing
- Command line
- Text mining
- Focused on better scientific programming practices, for writing software or analysis scripts
- Novice and more advanced levels
- Domain agnostic

Curriculum
- Command line
- Software development best practices in R, Python, or MATLAB
- Github for version control
The Software Carpentry Foundation runs the Instructor Training program that teaches volunteers teaching pedagogy and specific strategies for teaching computational skills in a workshop format.
Training is a missing piece between data collection & data-driven discovery
Results of the pre- and post-workshop surveys over the last 1.5 years (analyzed early 2017)
Q14 I can immediately apply what I learned at this workshop.

Answered: 592   Skipped: 85

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree
Q14 This workshop was worth my time.

Answered: 57  Skipped: 13

Q17 I would recommend this workshop to a friend or colleague.

Answered: 217  Skipped: 440
Q12 How would you rate your change in confidence in the tools that were covered during your Carpentry workshop(s) compared to before the workshop?

Answered: 321   Skipped: 46

- I'm more confident now.
- I'm equally confident now.
- I'm less confident now.

0%  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%
If you want to go fast, go alone.
If you want to go far, go together.
Get Involved

- Request a workshop
- Be a helper at a workshop
- Become a partner or affiliate and train local instructors
- Contribute to lesson development
Guiding the Carpentries

**Software Carpentry Steering Committee:**
Karin Lagenstrom (Norway)
Rayna Harris (UT Austin)
Susan McClatchy (Jackson Labs)
Kate Hertweck (UT Tyler)
Christina Koch (University of Wisconsin)
Mateusz Kuzak (eScience Netherlands)
Belinda Weaver (Australia)

**Data Carpentry Steering Committee:**
Karen Cranston (Duke / OpenTree of Life)
Hilmar Lapp (Duke)
Aleksandra Pawlik (New Zealand eScience)
Karthik Ram (rOpenSci / Berkeley Institute of Data Science Fellow)
Ethan White (University of Florida / Moore DDD Investigator)
Support

Data Carpentry

GORDON AND BETTY MOORE FOUNDATION

BEACON
SESYNC
NESCent
Data ONE
iDigBio
iPlant Collaborative

Software Carpentry

http://software-carpentry.org/scf/partners/

Jason Williams at iPlant