THE CARPENTRIES

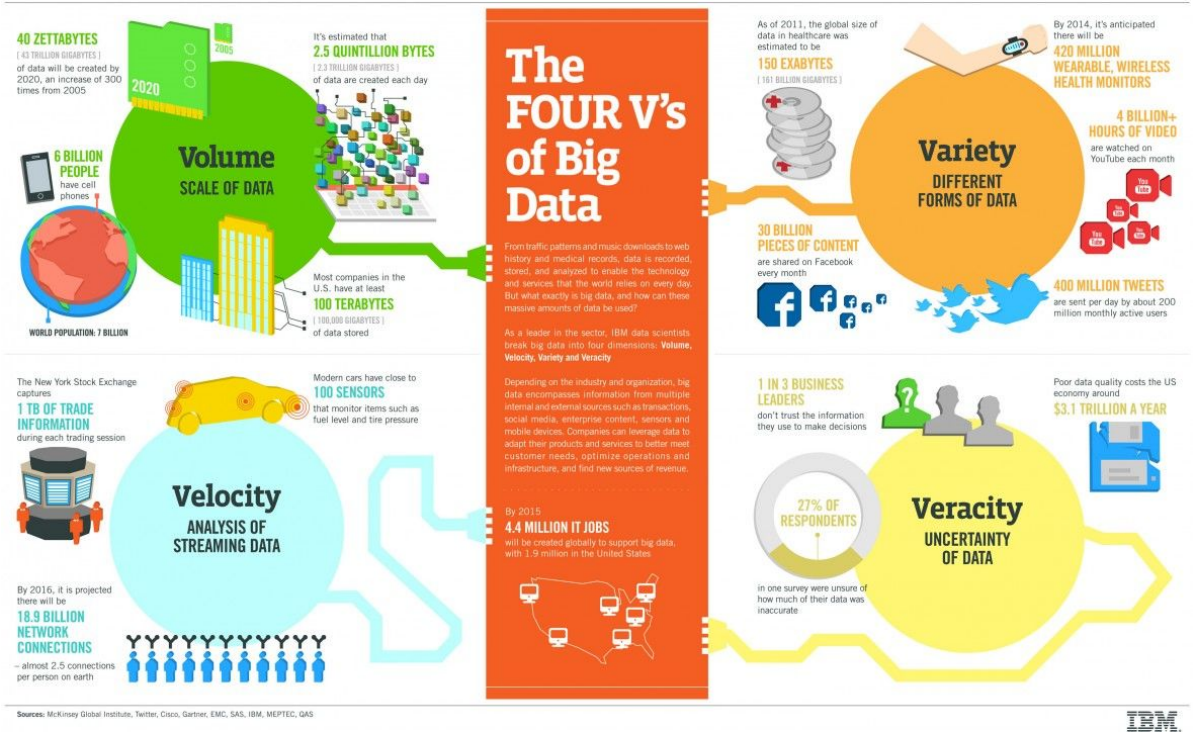
software carpentry

DATA CARPENTRY

library carpentry

Tracy K. Teal, PhD
Executive Director
@tracykteal @thecarpentries

Even in data-intensive fields of research we are generating even more and more types of data



This data has such potential.
Its limits are our capacity to work with this data.

- Data workflow knowledge
- Software development
- Efficient and effective software use

In particular, application development is critical both in broadening the use of petascale computing and in advancing to the exascale. Participants noted that “co-design” of architectures and algorithms (in which scientific problem requirements influence architecture design and technology constraints inform the formulation and design of algorithms and software) offers the opportunity to improve the effectiveness of both petascale and exascale systems. In addition, training programs are needed both to encourage use of resources at the lower levels of the “Branscomb pyramid” and to address the new operating models and different memory hierarchies expected for exascale systems.

NSF Advisory Committee for the Cyberinfrastructure Task Force on High Performance Computing, Final Report March 2011
https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_HPC.pdf

A variety of strategies will be needed to engage new communities. These include mentoring programs to encourage members of underrepresented groups to pursue careers in HPC; focused HPC outreach programs at the campus level, such as the NSF Campus Champions; and the development of tools and training to facilitate the use of HPC and computational techniques for R&E in additional disciplines. It should be noted that one community that previously had a large presence in HPC has diminished in activity: computational engineering.

NSF Advisory Committee for the Cyberinfrastructure Task Force on High Performance Computing, Final Report March 2011
https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_HPC.pdf

How do we scale data and software skills along with data production?

Building Skills and Community

- Creating training ‘in the gaps’ that is accessible, approachable, aligned and applicable
- Peer-led hands-on intensive workshops
- Volunteer instructors
- Open and collaborative lesson materials
- Creating and supporting community

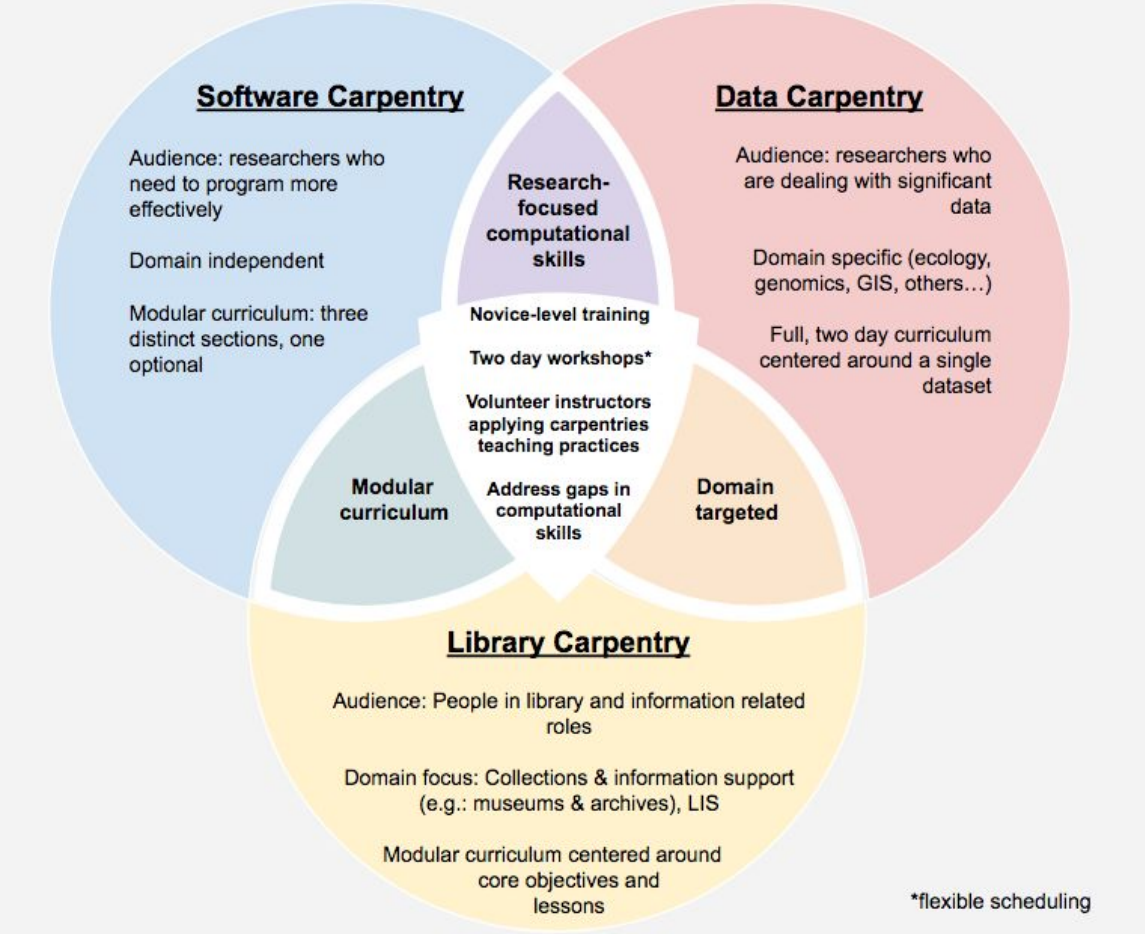
THE CARPENTRIES

Non-profit organization that:

- Trains people in software development and data science skills for more effective work and career development
- Builds community and local capacity for teaching and learning these skills and perspectives

Workshops

- 2-days, active learning
- Feedback to learners throughout the workshop
- Trained instructors
- Friendly learning environment


Software Carpentry
Audience: researchers who need to program more effectively
Domain independent
Modular curriculum: three distinct sections, one optional
Data Carpentry
Audience: researchers who are dealing with significant data
Domain specific (ecology, genomics, GIS, others…)
Full, two day curriculum centered around a single dataset
Research-focused computational skills
Novice-level training
Two day workshops*
Volunteer instructors applying carpentries teaching practices
Modular curriculum
Address gaps in computational skills
Domain targeted
Library Carpentry
Audience: People in library and information related roles
Domain focus: Collections & information support (e.g.: museums & archives), LIS
Modular curriculum centered around core objectives and lessons
*flexible scheduling

Data Carpentry Lessons

- Working effectively with data and includes domain-specific content

  Workflow for working with data
  - Data organization
  - Project organization
  - Data exploration and visualization
  - Automating workflows

Data Carpentry Lessons

- Domain content
  - **Ecology:** working with tabular data, ecological data
  - **Genomics:** cloud computing, genomic data organization, working with bioinformatics tools at the command line
  - **Geospatial:** organizing and working with geospatial data in R
  - **Social science**: tabular data with social science data
  - More in development

Software Carpentry Lessons

- Software development best practices
  - Command line
  - Version control with github
  - Programming in Python or R

Library Carpentry Core Objectives

Library Carpentry workshops teach people working in library- and information-related roles how to:

- Cut through the jargon terms and phrases of software development and data science and apply concepts from these fields in library tasks;
- Identify and use best practice in data structures;
- Learn how to programmatically transform and map data from one form to another;
- Work effectively with researchers, IT, and systems colleagues;
- Automate repetitive, error prone tasks.

Library Carpentry Core Lessons

- **Introduction to Data**
  An introduction to data structures, regular expressions, and computing terms (Jargon Busting & Pattern Matching)
- **The Unix Shell**
  An introduction to command line interfaces and task automation using the Unix shell (Text-based)
- **Introduction to Git**
  An introduction to version control using Git and GitHub for collaboration (GitHub Focus)
- **OpenRefine**
  An introduction to cleaning up and enhancing a dataset using OpenRefine (Journal Metadata Cleaning)

Workshop goals

- Teach skills
- Get people started and introduce them to what’s possible
- Build confidence in using these skills
- Encourage people to continue learning
- Positive learning experience

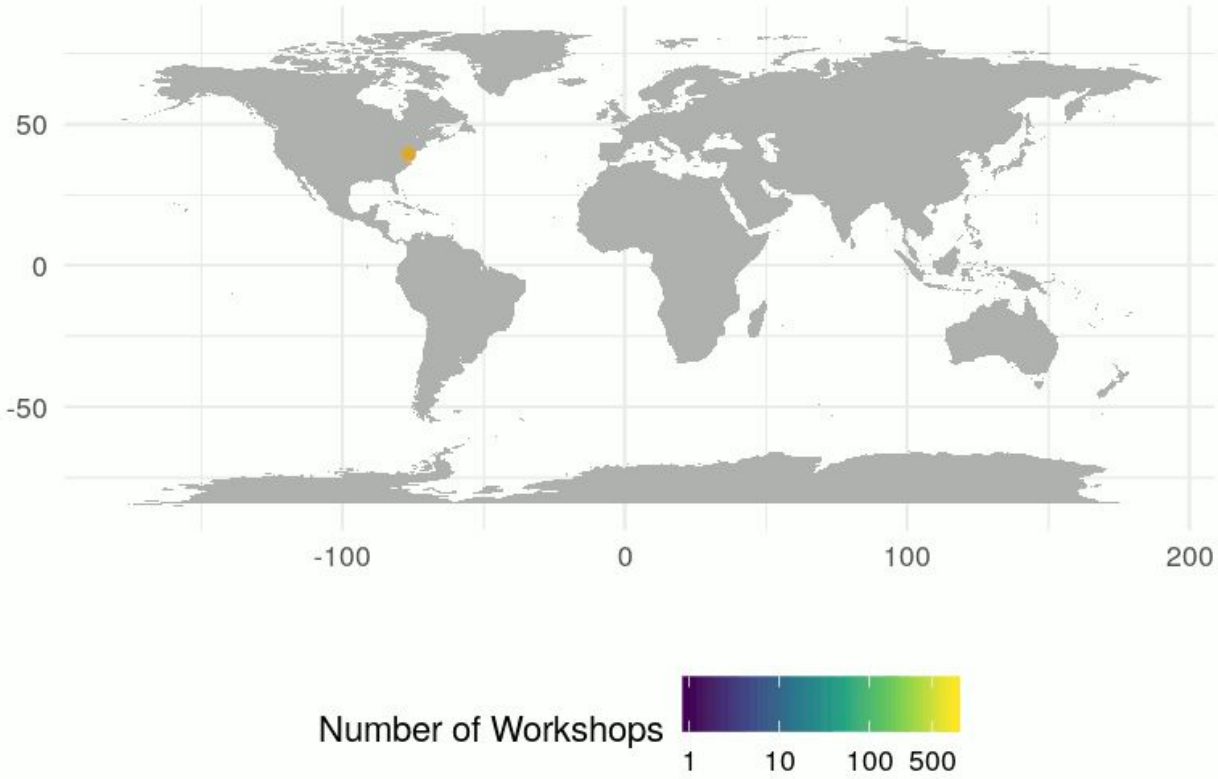
Our Workshops. Our learners.

1,640
instructors

39,000
learners

1,703
workshops

Workshops worldwide



Instructors

Instructor training program that teaches educational pedagogy. How to teach generally as well as for Carpentries workshops.

http://carpentries.github.io/instructor-training/

Over 1,600 volunteer instructors on 6 continents

Curriculum

- Open and collaboratively developed
- Continual improvement and up-to-date

Home | Code of Conduct | Setup | Episodes | Extras | License | Improve this page | Search...

Programming with Python

The best way to learn how to program is to do something useful, so this introduction to Python is built around a common scientific task: data analysis.

Our real goal isn't to teach you Python, but to teach you the basic concepts that all programming depends on. We use Python in our lessons because:

1. we have to use *something* for examples;
2. it's free, well-documented, and runs almost everywhere;
3. it has a large (and growing) user base among scientists; and
4. experience shows that it's easier for novices to pick up than most other languages.

But the two most important things are to use whatever language your colleagues are using, so that you can share your work with them easily, and to use that language *well*.

We are studying inflammation in patients who have been given a new treatment for arthritis, and need to analyze the first dozen data sets of their daily inflammation. The data sets are stored in comma-separated values (CSV) format: each row holds information for a single patient, and the columns represent successive days. The first few rows of our first file look like this:

```
0,0,1,3,1,2,4,7,8,3,3,3,10,5,7,4,7,7,12,18,6,13,11,11,7,7,4,6,8,8,4,4,5,7,3,4,2,3,0,0
0,1,2,1,2,1,3,2,2,6,10,11,5,9,4,4,7,16,8,6,18,4,12,5,12,7,11,5,11,3,3,5,4,4,5,5,1,1,0,1
```

Curriculum Development Process

In the process of developing infrastructure and guidelines to support more lesson development.

- Identifying needs for content
- Identifying learning goals and objectives
- Content development and assessment

Community

A group of people excited about software and data skills and about sharing them with others

- Mentoring program and instructor onboarding
- Discussion groups and community calls
- Email lists
- Teaching at other institutions

Outcomes

Short and long term surveys show that people are learning the skills, putting them into practice in their work and have more confidence in their ability to do computational work.

The tools I learned in my Carpentry workshop:

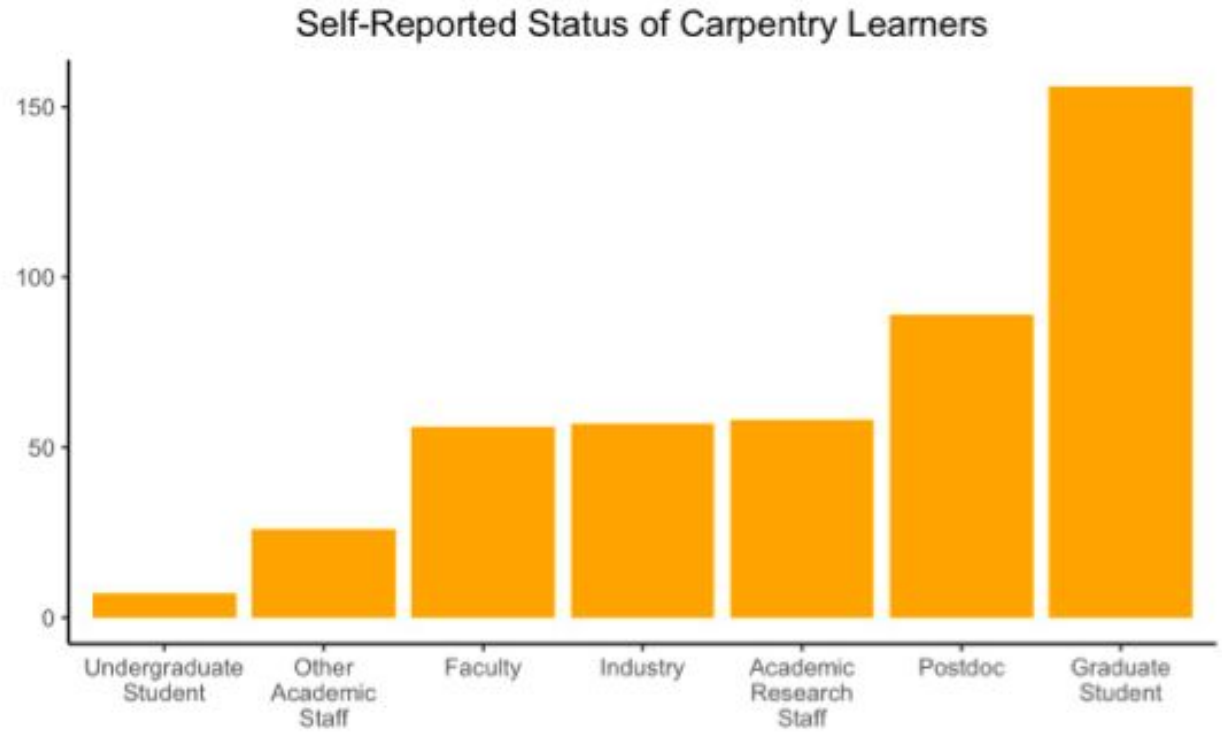
“helped me to reshape my workflow into a far more efficient and robust process.”

“are improving my ability to share data and code.”

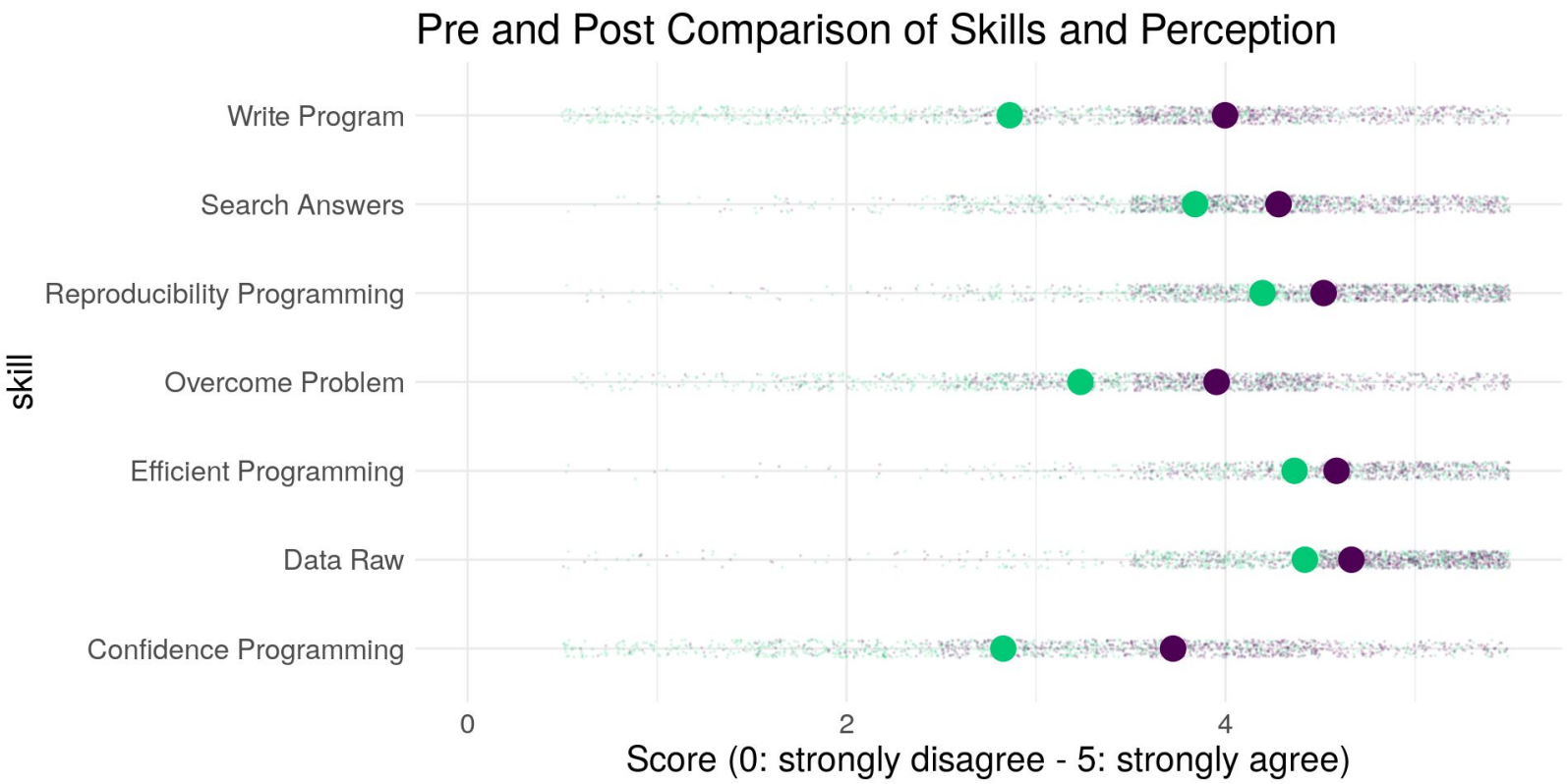
“helped facilitate my understanding of the problems and solutions to accessing and transforming data.”

“[are] useful tools for training my own team.”

Who takes workshops?

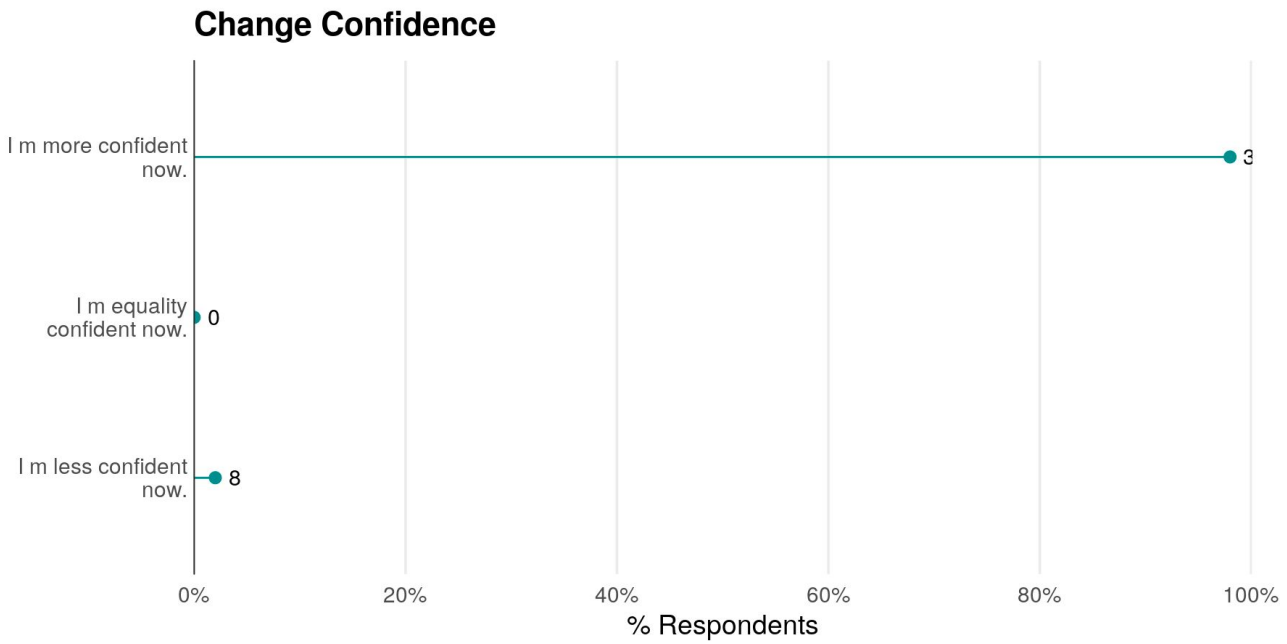


Confidence increases after just two days




examples
friendly
workshop
organized
analysis
research
instructor
questions
learning
programming
comfortable
informative
explanations
overview
knowledgeable
class
environment
website
excellent
online
exercises
presenters
sticky
explained
refine
knowledge
prior
helped
applicable
support
tools
software
relevant
introduction
basics
future
understanding
participants
beginners
live
materials
real
level
practice
intro
coding
lot
enjoyed
lessons
time
lots
trainers
2
extremely
concepts
basic
commands
skills
apply
experience
atmosphere
programs
instructions
covered
information
topics
learn
code
day
easy
taught
resources
students
instruction
helpful
follow
practical
understand
training
challenges
learned
provided
nice
sql
python
content
data
material
openrefine
patient
step
management
teaching
hands
helpers
teachers
interactive
people
pace

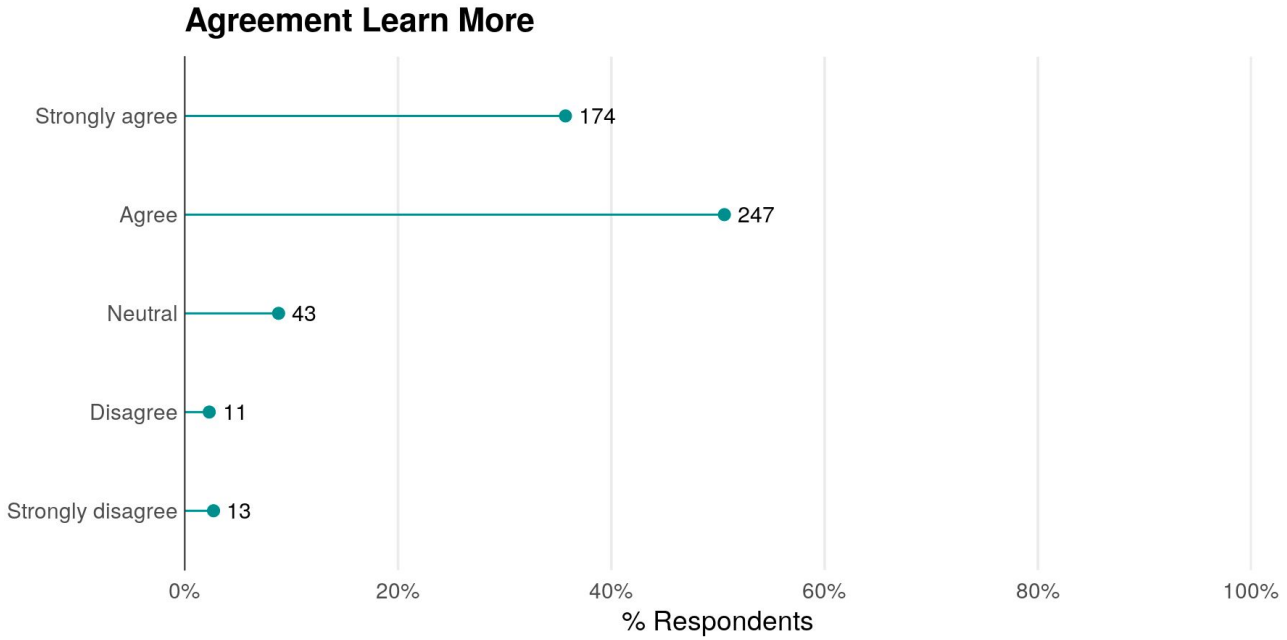
Confidence persists long term



Numbers of answers reported on the graph (n = 404).

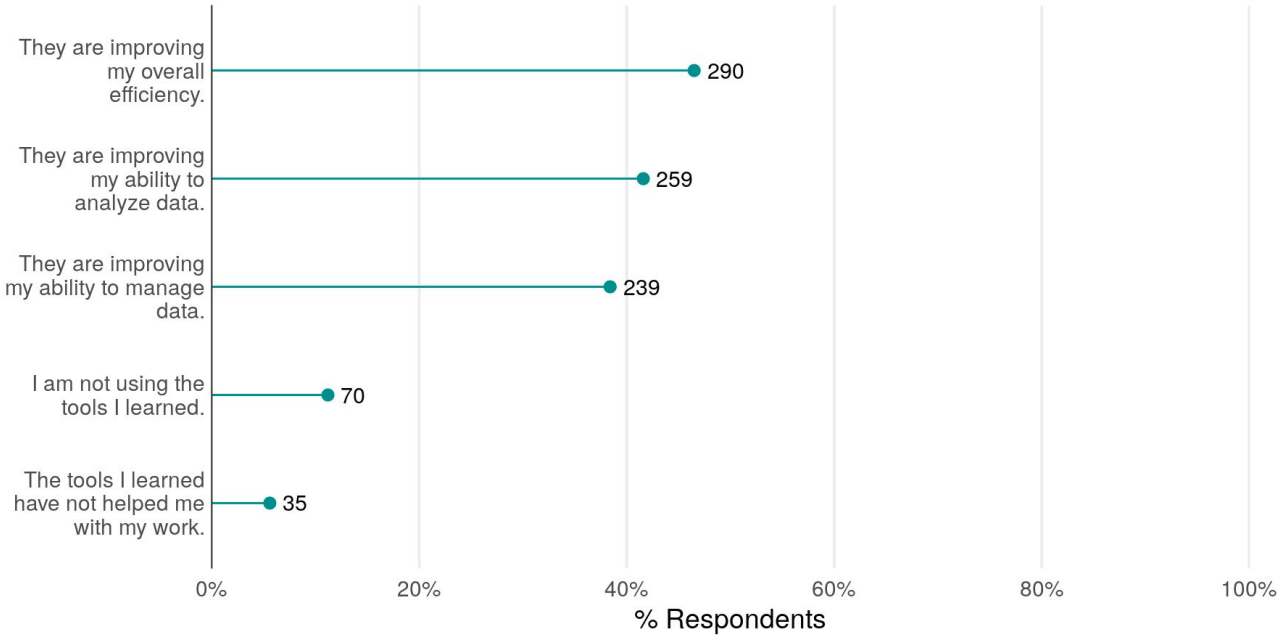
Continued learning

People continue to learn after workshops



Numbers of answers reported on the graph (n = 488).

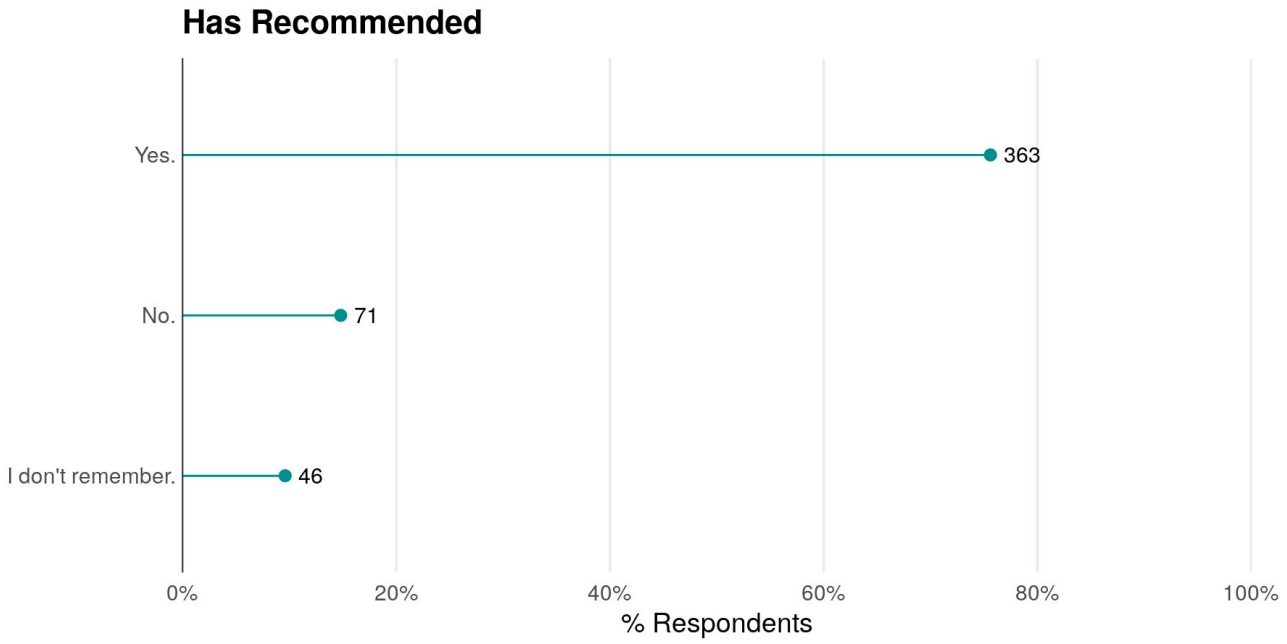
How are these skills affecting your work



Numbers of answers reported on the graph (n = 893).


version
workflow
workshop
python
understanding
improved
started
carpentry
line
learned
teaching
command
shell
language
scripting
daily
2
analysis
collaboration
workflows
sql
training
basis
collaborate
lab
unix
community
text
code
automate
confidence
github
project
1
codes
tools
implemented
skills
increased
time
account
coding
students
scripts
git
team
software
programming
thesis
learn
familiarized
https
learning
control
introduced
github.com
improve
knowledge
data
research
comfortable

People strongly recommend workshops to others



Numbers of answers reported on the graph (n = 480).


If you want to go fast, go alone.
If you want to go far, go together.

Support

GORDON AND BETTY MOORE FOUNDATION

Alfred P. Sloan FOUNDATION

INSTITUTE of Museum and Library SERVICES

Software Sustainability Institute