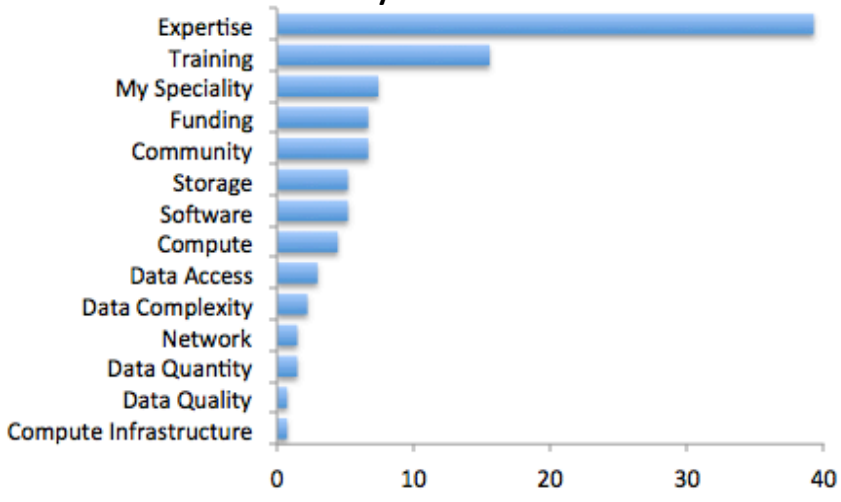
Using the Carpentries

Tim Dennis, UCLA @jt14den
Juliane Schneider, Harvard Catalyst @JulianeS
Library Carpentry: @libcarpentry

Researchers view the major limiting factor in research progress as a lack of expertise in how to handle and analyze data

Survey by Bioinformatics Resource Australia – EMBL

Biggest Bioinformatics Difficulty



Most useful thing BRAEMBL could do

http://braembl.org.au/news/braembl-community-survey-report-2013

In a 2016 survey of 704 NSF Biological Sciences Directorate principle investigators… **nearly 90% indicated they are currently or will soon be analyzing large data sets**. …. the **most pressing unmet needs are training in data integration, data management**, and scaling analyses for HPC … This portends a **growing data knowledge gap** in biology and challenges institutions and funding agencies to **redouble their support for computational training** in biology.

Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators https://doi.org/10.1101/108555

New knowledge and skills will be needed to make effective use of new system architectures and software. “Hybrid” disciplines such as **computational science** and **data science** and interdisciplinary teams may come to **play an increasingly important role**. Keeping abreast of a **rapidly evolving suite of relevant technologies** is challenging for many computer science programs ...

Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: Interim Report

How do we bring effective, ongoing instruction to our institutions?

New instructional needs + training ourselves = finding resources for both

If we provide effective data practice and computational instruction, better datasets and better research shall result

Software Carpentry

Software development best practices

Domain agnostic

- Command line
- Version control with git
- Programming in Python or R

Data Carpentry

Working effectively with data
includes domain-specific content

- Data organization
- Data cleaning
- Data analysis and visualization in R or Python

Library Carpentry

Working effectively with data using best practices

- Basic computation skills
- Versioning and collaboration through Github
- Data analysis and cleanup through OpenRefine

Overall 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

Curriculum and Communication

- Open and collaboratively developed
- Continually improved and up-to-date
- Uses Github, Gitter and Etherpads

The Carpentries at UCSD

A Use Case

- Library offered first SWC workshop in Oct. 2015
- Began offering stand-alone R workshops around same time
- Partnered with SWC in Spring 2016
- Held week-long Library Carpentry workshop July 2016

The Carpentries at UCSD

What happened?

- Carpentry lessons formed the core of curriculum we called “Data Workshops”
- 2015-17 we offered 20+ Data Workshops
- Over 850 learners came from 12 different academic departments
- Offered 4 2-day and 18 one-off workshops
- One week-long Library Carpentry workshop

Software Carpentry

R, Python, Bash & Git

3 workshops (Python)

Worked with SIO to integrate SWC workshop into a introductory progression in Genomics

Lessons can be used for one-off stand-alone courses

Data Carpentry

R for Genomics

Worked with Bioinformatics Librarian to offer stand-alone R workshop

Used lessons in Medical Informatics course instruction in conjunction with systematic review

Library Carpentry

Intro to data and jargon busting; Regular Expressions; Bash/ Shell; Git/Github; OpenRefine

Big leap to command-line thinking

Text manipulation and data cleanup

OpenRefine was a big hit

Adoption of GitHub by digital object metadata group

Gave librarians a better sense of tools used in computational research

Reusing Carpentry Lessons in the Curriculum

Design 9 week course with UCSD Global Policy School

3 weeks each of R, Python and SQL/Data Management

Reworked lesson challenges as quizzes/ assignments

Carpentry in the Curriculum

Outcomes

This course has been a huge success, is being taught for a second time this year [2017], and had more than 60 students receive certifications last year. I can say with confidence … ***this course has singlehandedly shifted the skills and consciousness of our students in important ways****; introducing R more deeply at a time it was little-used, spearheading a push towards more open-sourced software, and* ***most importantly bringing the best traditions of library data science to a large group of MA students shortly to enter the professional workforce****.*

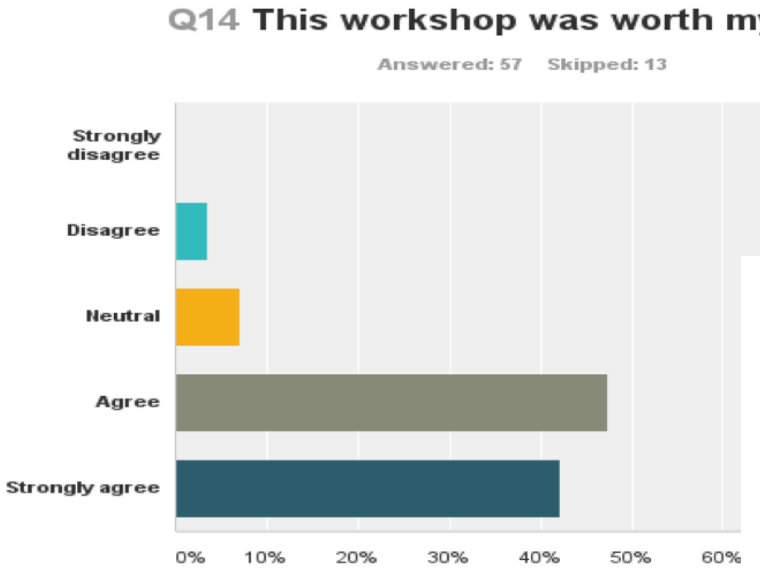
-- Craig McIntosh, Professor; Co-director, Policy Design and Evaluation Lab, UCS School of Global Policy and Strategy

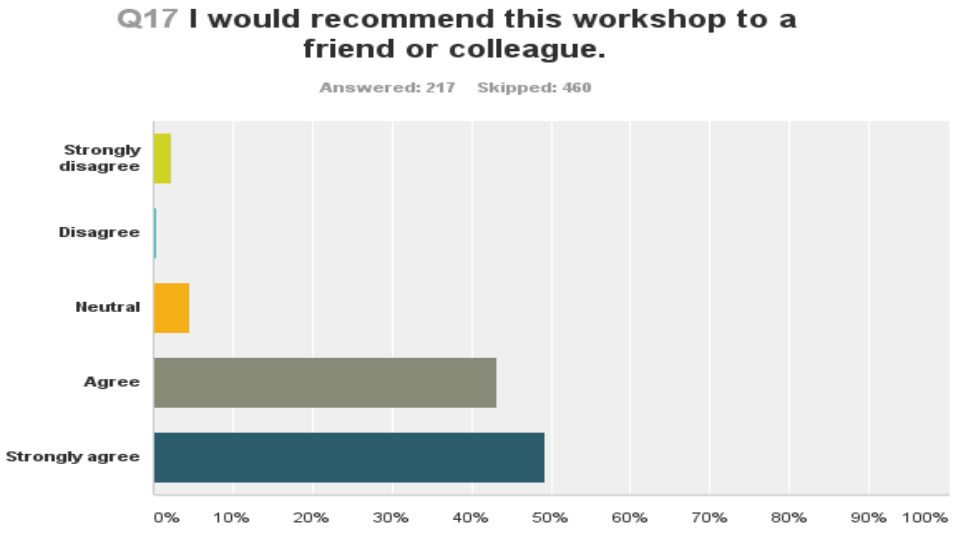
Community

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

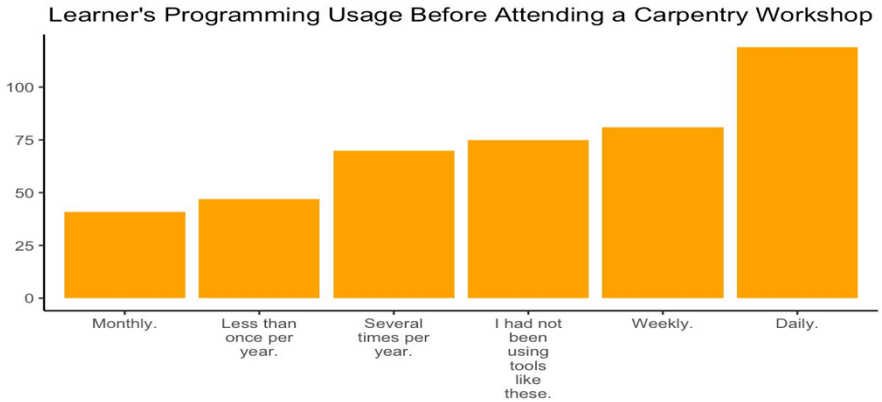
- Mentoring program and instructor onboarding
- Discussion groups and community calls
- Email lists
- Teaching at other institutions

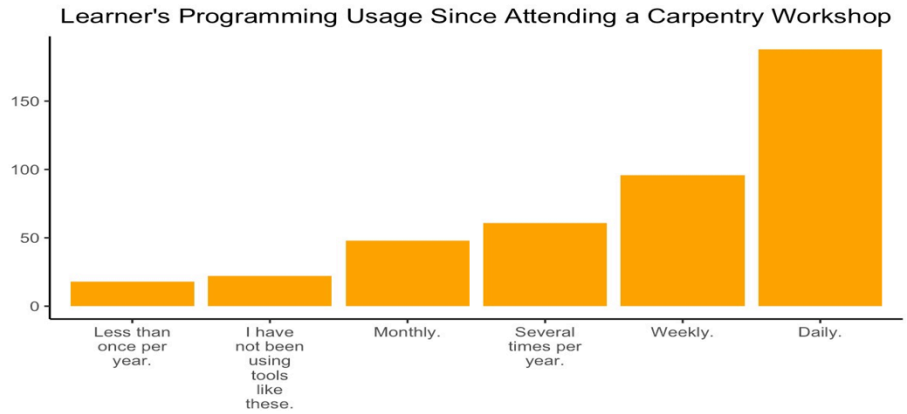
People like the workshops



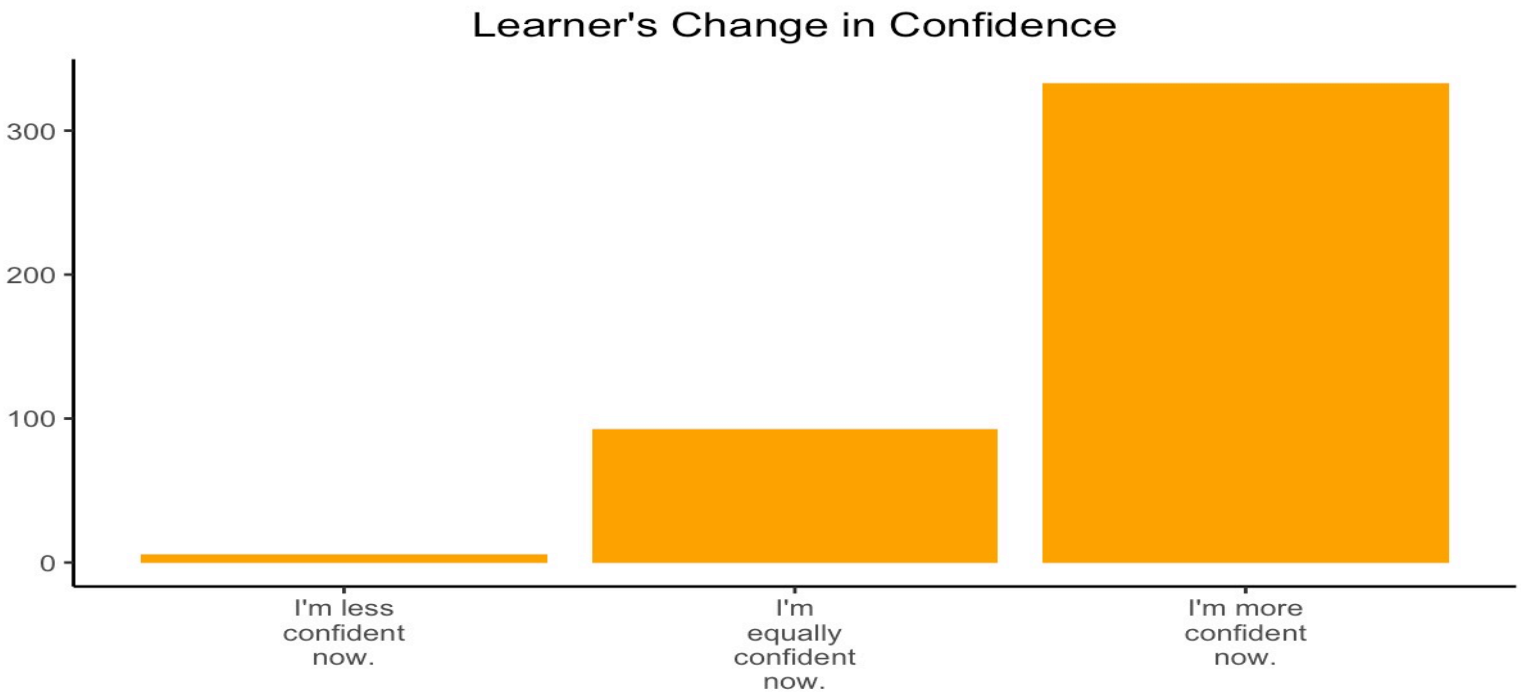


Workshop goals
skills, confidence, practical use



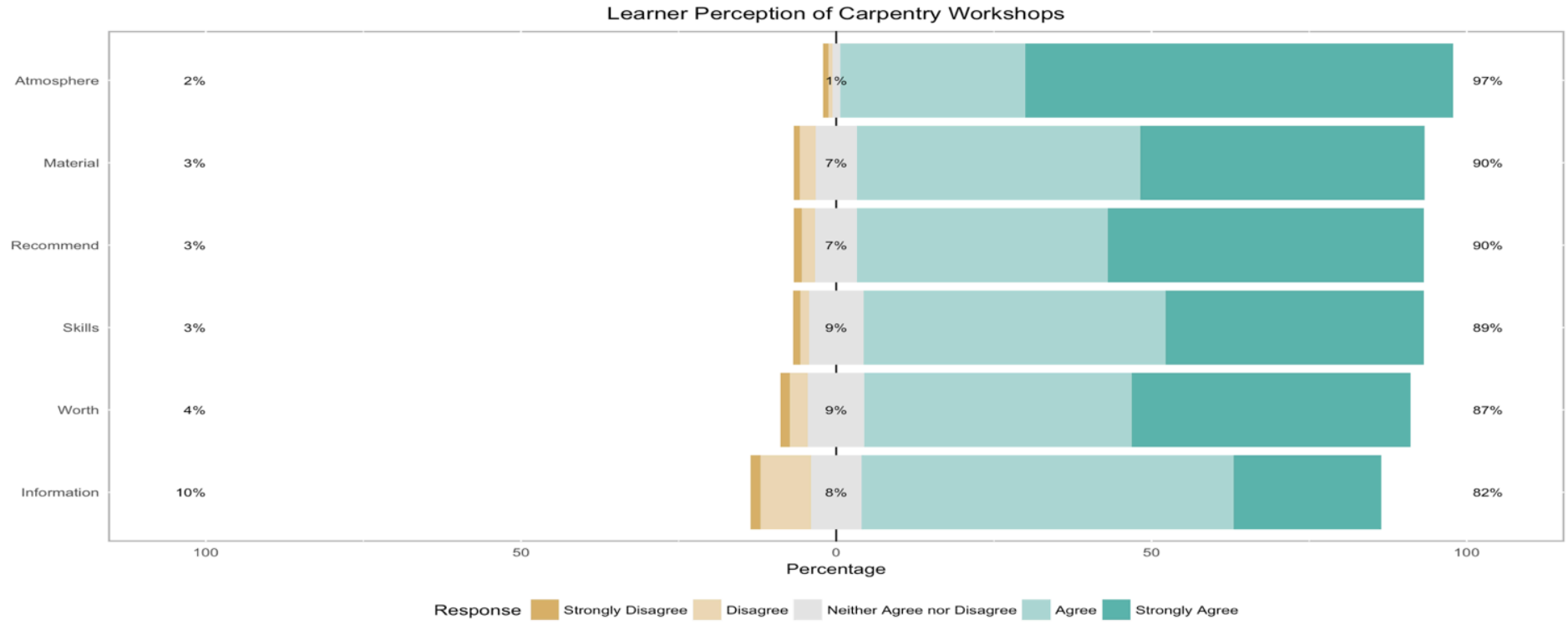


Confidence


Learner's Change in Confidence
300
200
100
0
I'm less confident now.
I'm equally confident now.
I'm more confident now.

This is the most important takeaway that everyone has to remember.

People have more confidence and continue learning 6 months or more after workshops



Training is a missing piece between data collection & data-driven discovery

All The Beautiful Links

Software/Data Carpentry:
https://software-carpentry.org/
Sofware Carpentry Pad of Pads:
https://software-carpentry.org/pad-of-pads
Library Carpentry:
http://librarycarpentry.github.io/
Library Carpentry Sprint Etherpad:
http://pad.software-carpentry.org/lc2017
Library Carpentry Gitter:
https://gitter.im/weaverbel/LibraryCarpentry