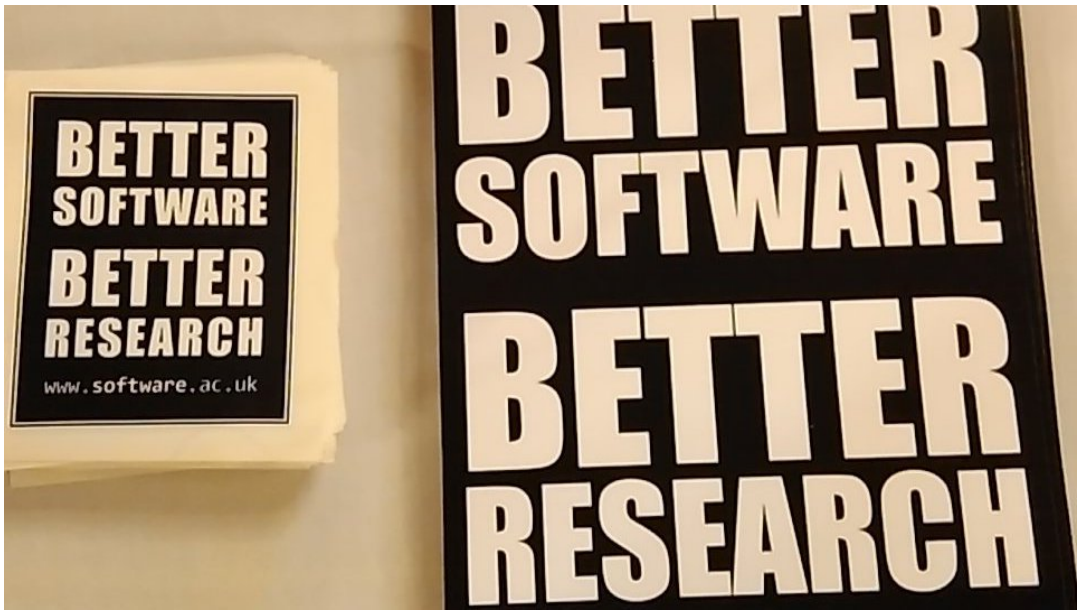
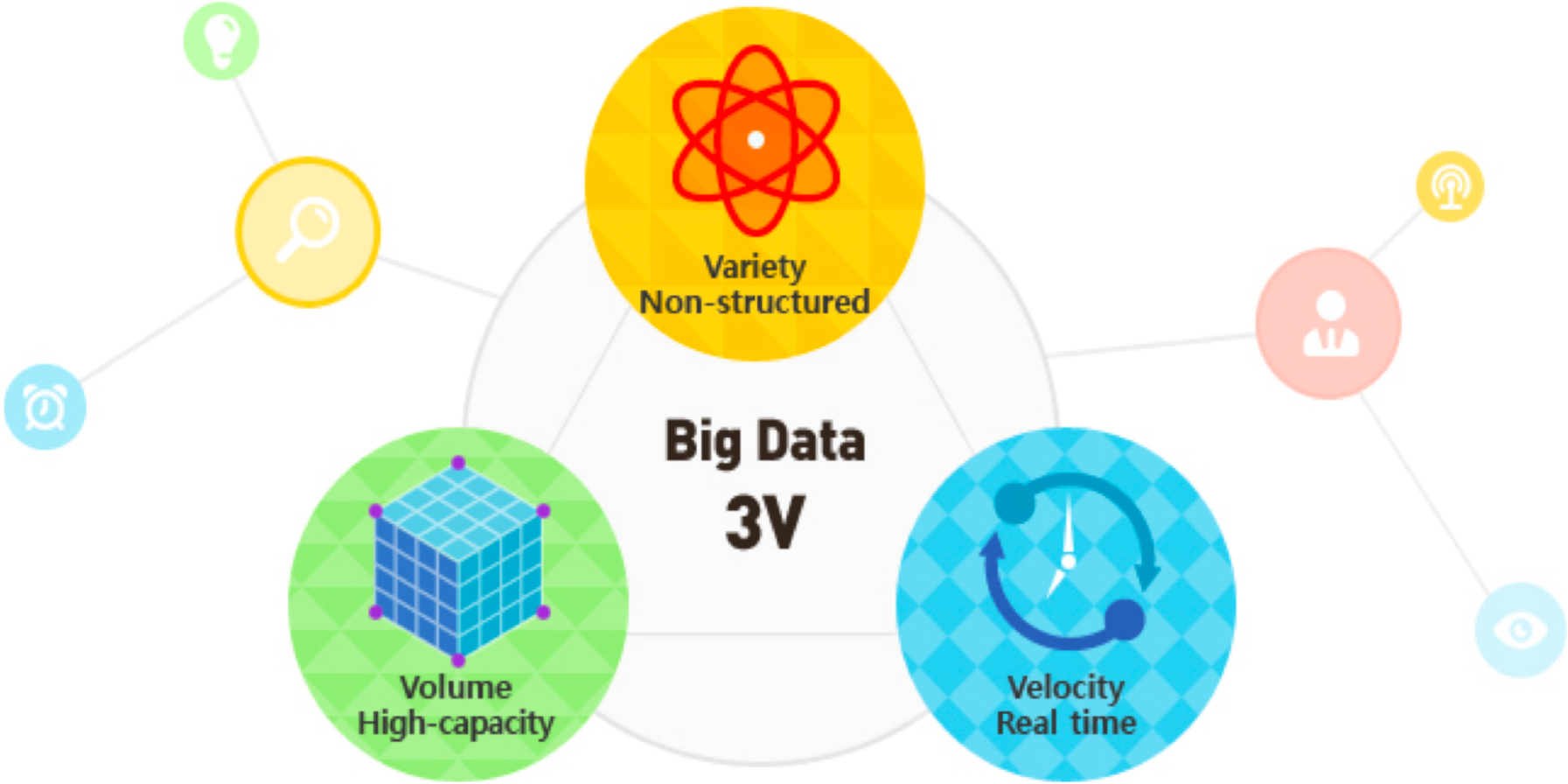
Data skills and software training to enable data-driven-discovery

Tracy K. Teal, PhD
Data Carpentry, Executive Director
@tracykteal, @datacarpentry
Find these slides on GitHub: https://github.com/tracykteal/msdse-presentation/


BETTER
SOFTWARE
BETTER
RESEARCH
www.software.ac.uk
BETTER
SOFTWARE
BETTER
RESEARCH

Happy Coder. Better Research.


Variety
Non-structured
Big Data
3V
Volume
High-capacity
Velocity
Real time

Software and tools allow us to turn data into information.

People turn information into knowledge.

By making data accessible and putting the data skills and the perspectives in the hands of all researchers, we allow them to answer their own questions and capture their passion and expert knowledge.

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

Training

Researchers want training

Most useful thing Bioinformatics Resource Australia can do

BRAEMBL community survey report

How do we provide training?

- Training in the gaps
- Scaling curriculum and instructors
- Friendly learning environment

Training in the Gaps

Training active researchers

Active researchers and employees are learning these skills "on the job".

Need to develop and deliver training that fits their time and needs.

- Training that is immediate, accessible, appropriate for their level and relevant to their domain.
- Include not only technical skills, but also ways of thinking about data and knowing what's possible
- Opportunity for deliberate practice, hands-on training with feedback during learning
- Researchers need to build confidence and the belief that they are capable of computational work, self-efficacy

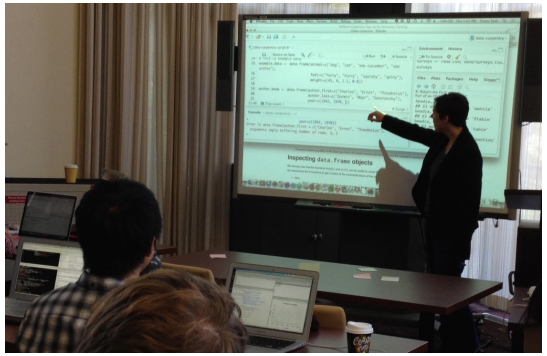
Software Carpentry and Data Carpentry

Community organizations training researchers in best practices for data analysis and software development to make research more effective and reproducible.

- Core skills for effective research computing
- Two-day hands-on workshops
- Collaboratively developed, openly licensed lesson materials
- Over 700 trained volunteer instructors on 6 continents

How we teach: Hands on intensive workshops

- Short: Two days
- Impactful: Focused time
- Convenient: Held at the university or organization
- Interactive: Hands-on teaching and exercises
- Immediate feedback: sticky notes & minute cards
- Qualified instructors
- Shared learning and a friendly learning environment



What we teach

Foundational skills needed to effectively conduct computational research.

Software Carpentry Domain-agnostic and focused on better coding practices.

Data Carpentry
Domain specific and focused on how to work effectively with data.

Lesson	Site	Repository	Reference	Maintainer(s)
The Unix Shell				Gabriel Devenyi, Ashwin Srinath
Version Control with Git				Ivan Gonzalez, Daisie Huang
Version Control with Mercurial				Doug Latornell
Using Databases and SQL				Abigail Cabunoc Mayes, Sheldon McKay
Programming with Python				Trevor Bekolay, Valentina Staneva
Programming with R				Daniel Chen, Harriet Dashnow
R for Reproducible Scientific Analysis				Thomas Wright, Naupaka Zimmerman
Programming with MATLAB				Isabell Kiral-Kornek, Ashwin Srinath
Automation and Make				Gerard Capes
Instructor Training				Erin Becker, Christina Koch

Lessons

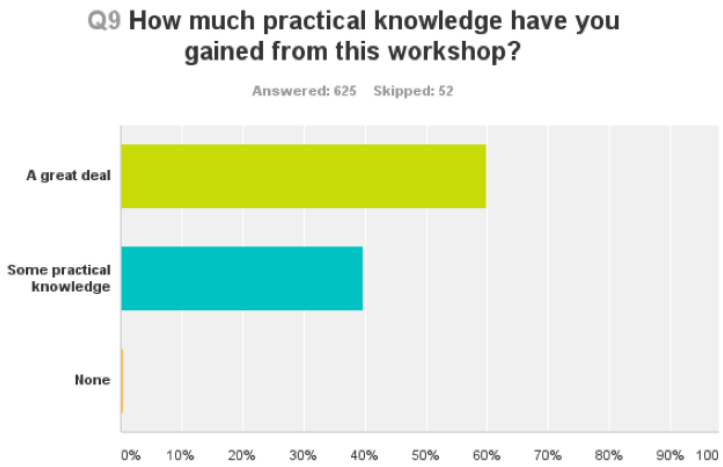
Lesson	Site	Repository	Instructor Guide	Maintainer(s)
Data Organization in Spreadsheets				Christie Bahlai, Tracy Teal
Data Cleaning with OpenRefine				Deborah Paul, Cam Macdonell
Data Management with SQL				Paula Andrea Martinez, Timothée Poisot
Data Analysis and Visualization in R				François Michonneau, Auriel Fournier
Data Analysis and Visualization in Python				John Gosset, April Wright, Mateusz Kuzak

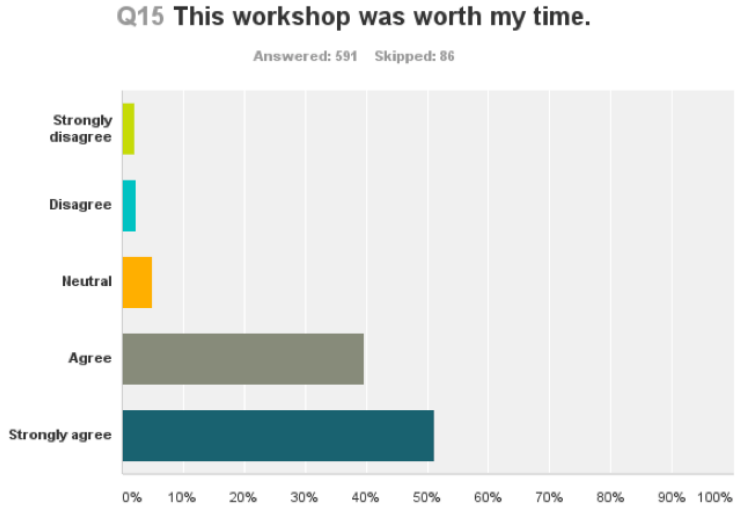
People are learning the skills

Table 8: Compared to before the workshop I have a better understanding of how to…

Item	n	Mean	Mode
Effectively organize data in spreadsheets	381	4.18	4
Use OpenRefine for data cleaning	311	4.61	5
Import a file into Python and work with the data	143	4.92	4
Import a file into R and work with the data	374	4.49	5
Do initial visualizations in Python	**145**	**4.90**	**3**
Do initial visualizations in R	369	4.47	5
Construct a SQL query statement	273	4.61	5
Use the command line	326	4.47	5

Skills are practical & workshops worth their time





Scaling curriculum and instructors

Sharing curriculum and creating instructors

- Collaboratively developed, openly accessible lessons
- http://www.software-carpentry.org/lessons
- http://www.datacarpentry.org/lessons
- Instructor training program

Since 2015

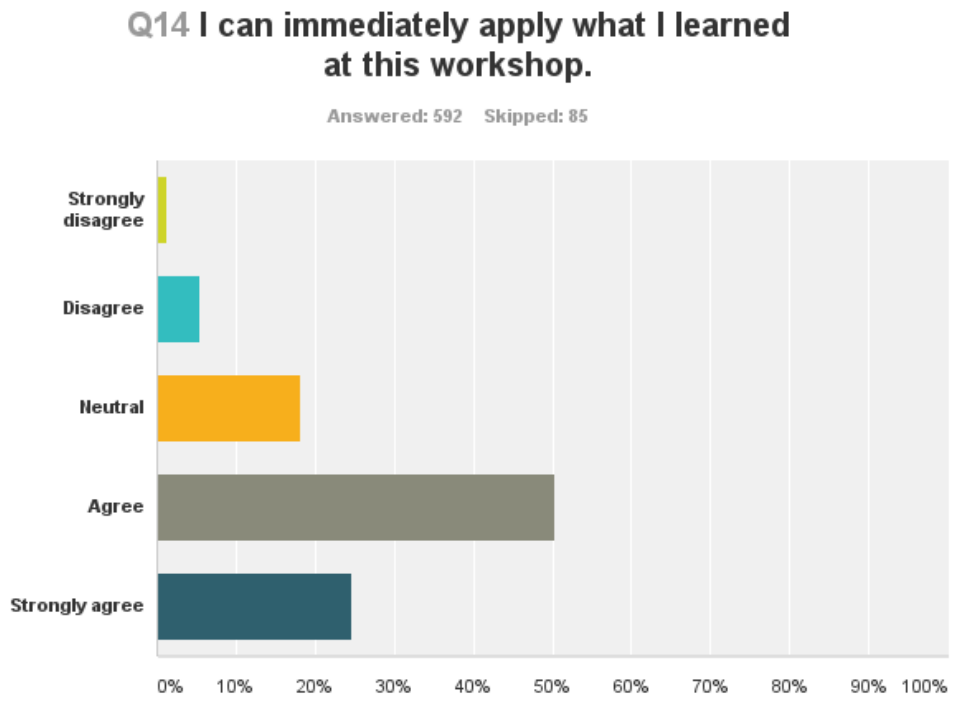
- 690 workshops
- 14,642 learners
- 6 continents

Friendly learning environment

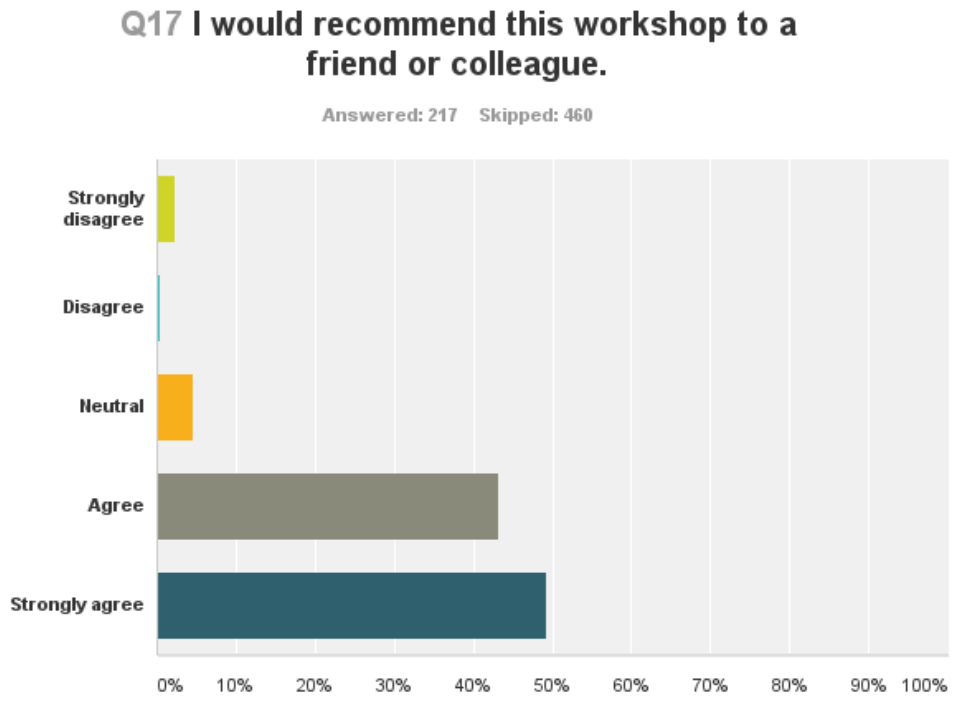
Need confidence to continue learning

- Meeting learners where they are
- Code of conduct

Learners feel they can apply what they learned



Would recommend the workshop to a colleague or friend




good
data
instructors
great
tools
introduction
workshop
people
different
help
helpers
presented
useful
helpful
teachers
hands
clear
knowledge
organization
management
programs
software
course
materials
easy
working
excellent
learning
teaching
notes
basic
new
practical
time
problems
analysis
green
examples
beginners
exercises
interactive
enthusiastic
python
sticky
format
presentation
worked
knowledgable
atmosphere
experience
sql
instructor
helped
etherpad
access
material
exposure
approach
taught
level
learned
information
explained
skills
organized
content
available
idea
following
refine
challenges
research
explanations
basics
everything
hands-on
covered
knowledgeable
someone
one
lots
nice
work

How can you or your organization be involved

- Request a workshop
- Become a Partner and build local training capacity
- Become an instructor or help at a workshop
- Contribute to lessons
- Join our 'announce' list to be a part of the community
- Be an advocate for training initiatives & opportunities

Acknowledgements

- Over 700 volunteers worldwide that teach and develop lessons
- Greg Wilson, who founded Software Carpentry
- The Steering Committees of Software and Data Carpentry (Karthik Ram and Ethan White)
- Software and Data Carpentry staff: Jonah Duckles, Greg Wilson, Erin Becker, Maneesha Sane and Kari Jordan

Acknowledgements

- Gordon and Betty Moore Foundation
- National Science Foundation BIO Centers: iDigBio, CyVerse, NESCent, SESYNC, BEACON, NEON

Happy Coders. Better Software.
Better Research.

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY