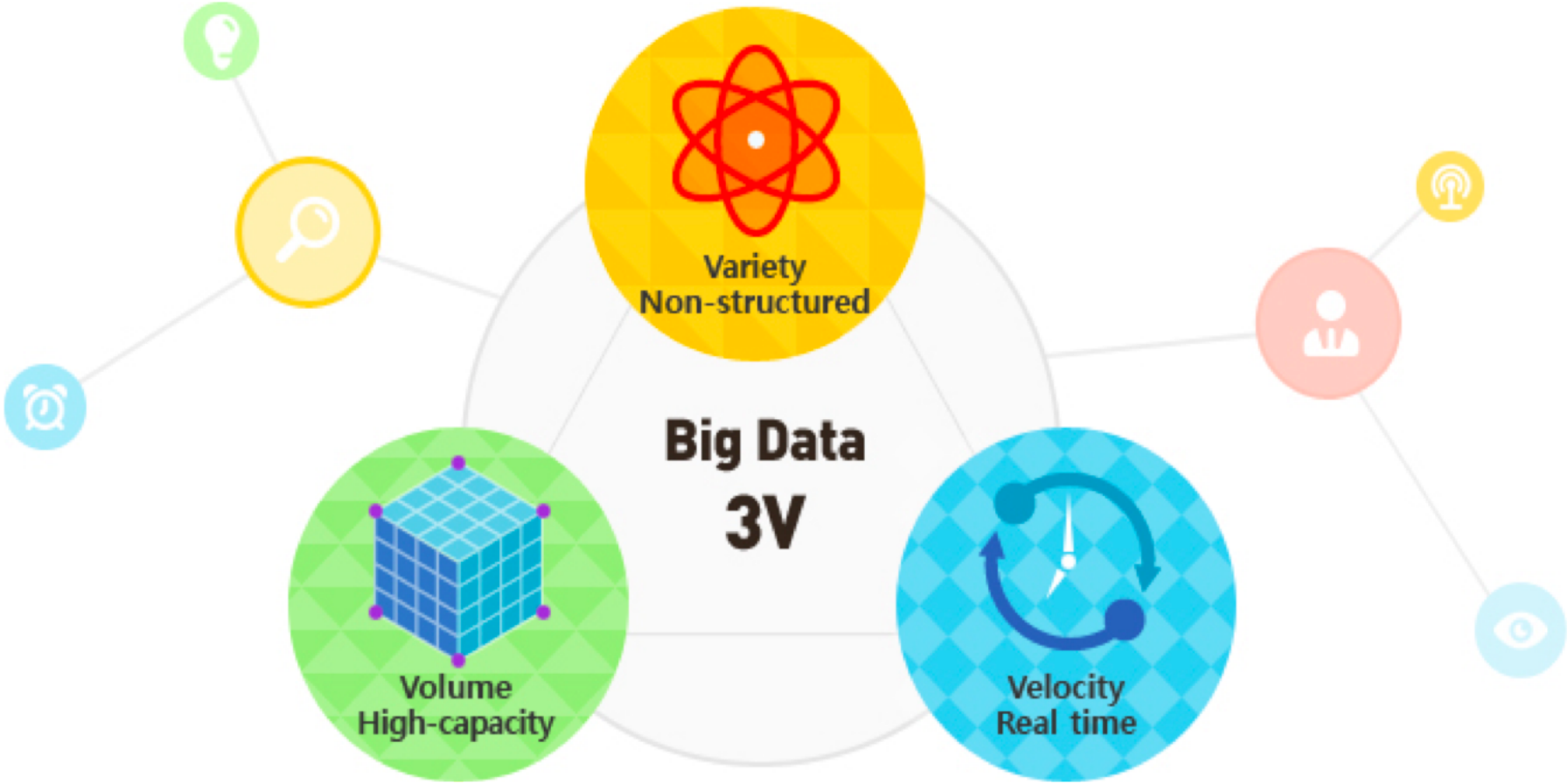
Democratizing data skills to advance data-driven-discovery

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


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.

Unleash the potential of data by empowering people

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

Researchers want these skills

Most useful thing Bioinformatics Resource Australia can do

BRAEMBL community survey report

Training in the Gaps: 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

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

Data Carpentry

Data Carpentry teaches domain-specific two-day, **hands-on workshops** on the foundational concepts, skills and tools for working more effectively with data.

No prior computational experience is required to attend workshops.

Workshops are taught by trained volunteer instrductors.

Lesson materials are collaboratively and openly developed.

Workshop goals

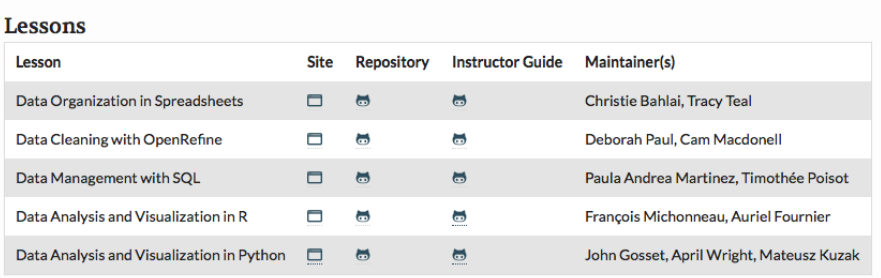
- Teach skills
- Build confidence
- Change perspectives

What we teach

Focused on data - teaches how to manage and analyze data in an effective and reproducible way.

Biology

- Data organization with spreadsheets
- OpenRefine for data cleaning
- SQL for data management
- R or Python for data analysis and visualization



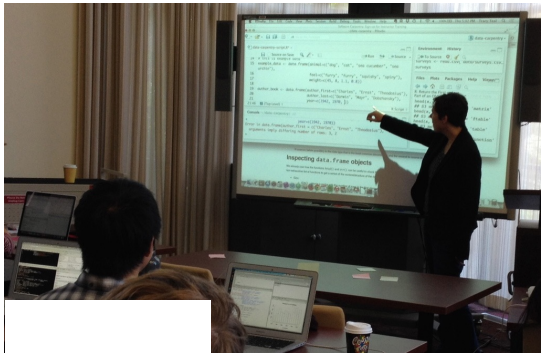
Lessons

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

Genomic and Geospatial data lessons cover topics relevant to those data types. Working on social sciences, library and image analysis.

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



Instructors

Outcomes

- Learner outcomes
- Instructor outcomes
- Community!

Workshop goals

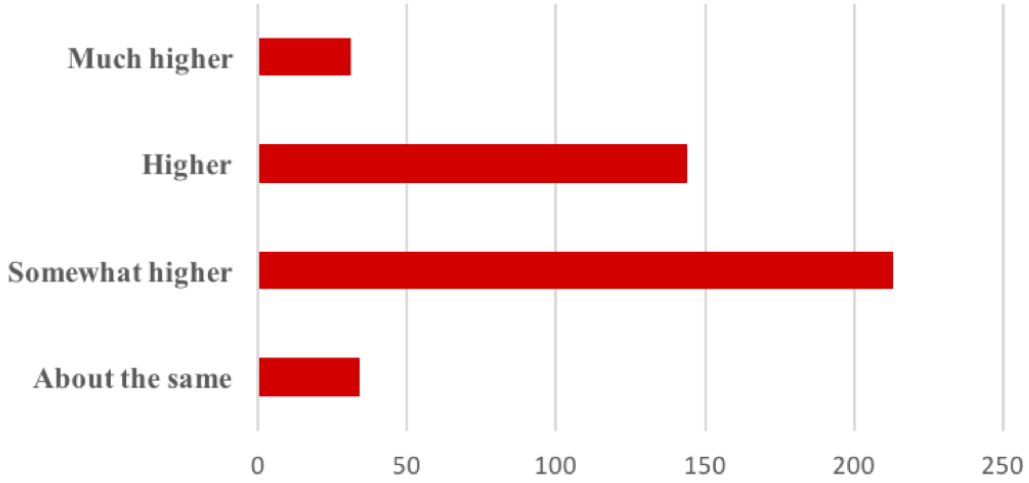
- Teach skills
- Build confidence
- Change perspectives

Increased skill levels reported

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

Figure 10: Level of data management and analysis skills following to the workshop



Increased confidence in ability to use skills

Figure 6: I can immediately apply what I learned at the workshop.
(n = 421)

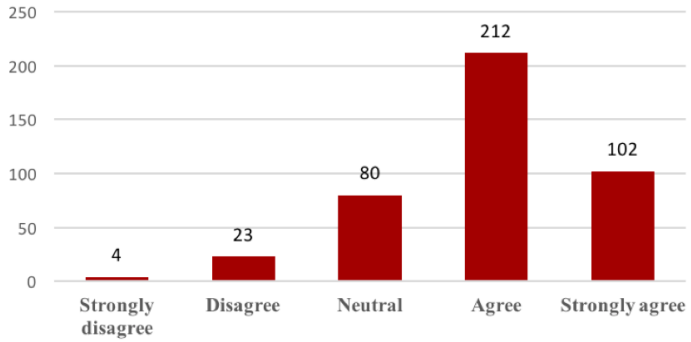
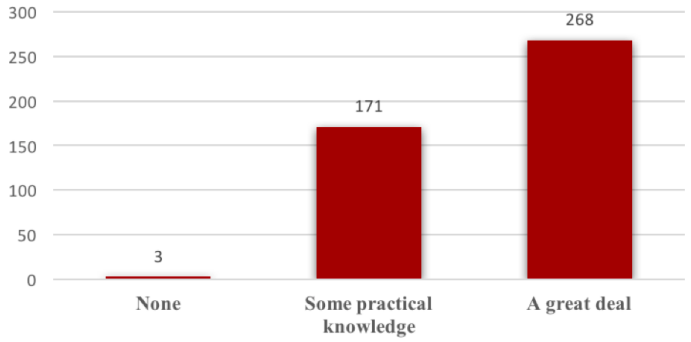


Figure 8: How much practical knowledge have you gained from this workshop?
(n = 478)



Change in perspective about use of skills for reproducible research

Table 9: Analysis of Learners' Research Computing Attitudes ***(Pre-Workshop)***

Item	n	Mean	Mode
Data organization is a fundamental component of effective and reproducible research.	923	4.47	4
Using a scripting language like R or Python can ultimately improve my analysis efficiency.	920	4.19	4
Using R or Python makes analyses easier to reproduce.	918	3.95	3
A value of using SQL, R or Python is that underlying data cannot accidentally be changed.	912	3.50	3

Table 10: Analysis of Learners' Research Computing Attitudes ***(Post-Workshop)***

Item	n	Mean	Mode
Data organization is a fundamental component of effective and reproducible research.	422	4.64	5
Using a scripting language like R or Python can ultimately improve my analysis efficiency.	422	4.43	5
Using R or Python makes analyses easier to reproduce.	420	4.39	5
A value of using SQL, R or Python is that underlying data cannot accidentally be changed.	419	4.06	5

Figure 5: The workshop was worth my time
(n = 419)

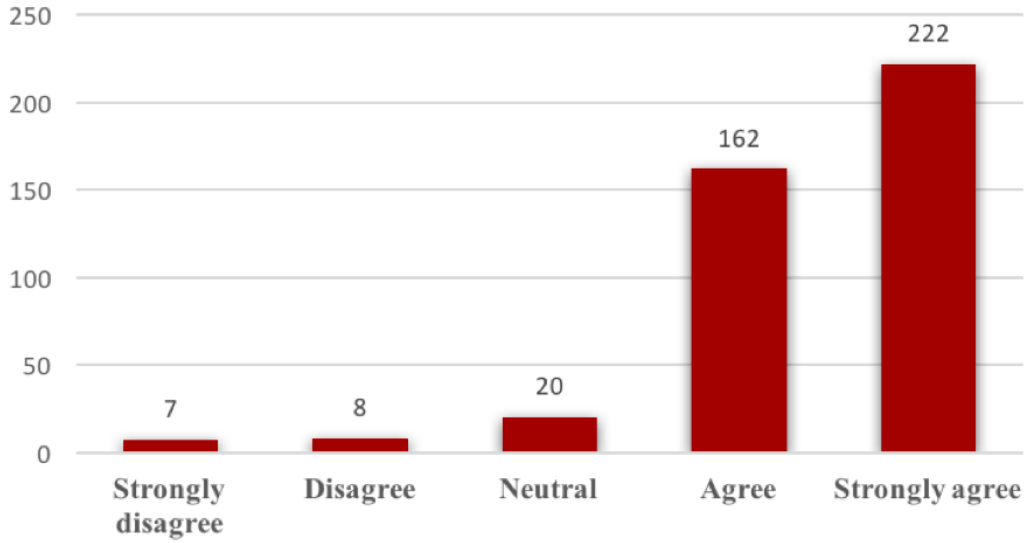
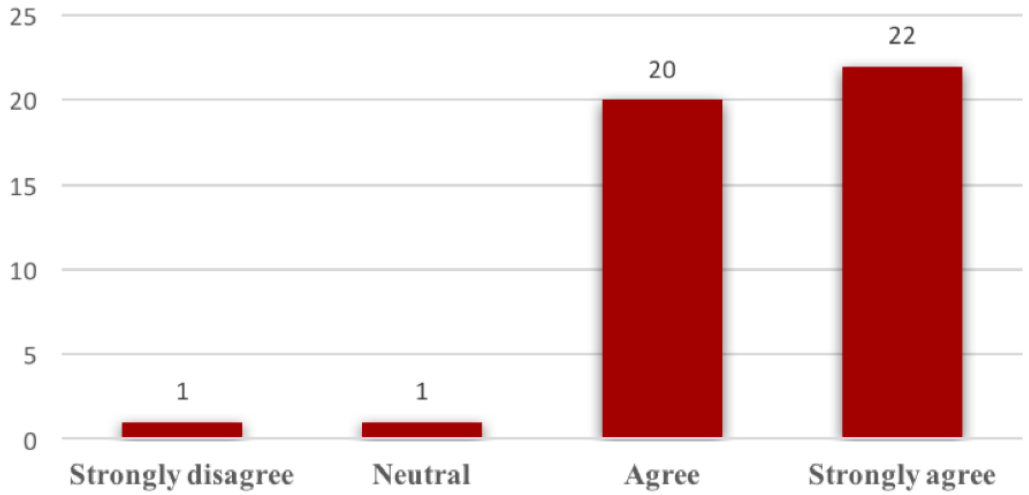


Figure 3: I would recommend this workshop to a colleague (n = 44)




good
instructors
data
great
introduction
tools
workshop
people
different
help
helpers
helpful
useful
presented
clear
hands
teachers
programs
management
organization
knowledge
software
materials
easy
course
working
teaching
learning
excellent
practical
notes
basic
new
time
problems
analysis
green
interactive
enthusiastic
python
sticky
format
presentation
worked
atmosphere
knowledgable
experience
examples
beginners
exercises
sql
explained
skills
instructor
helped
etherpad
access
organized
content
idea
available
following
refine
challenges
research
explanations
everything
basics
hands-on
taught
material
approach
exposure
information
level
learned
covered
knowledgeable
someone
one
nice
lots
practice
work
programming
effective
online
structure
particularly
explanation
teacher
experienced
learn
easily
ability
opportunity
pretty
unfamiliar
understand
possible
food
see
packages
assistants
tool
powerful
liked
personal
highly
mostly
skill
grasp
sure
real
environment
made
take
go
taster
helping
nature
needs
beginner
order
programmes
never
best

Instructor outcomes

- People learn new technical skills
- They have a community - no more 'lone informatician'
- People become better communicators
- Gain value from giving back and empowering people with the skills they have learned are valuable

Community!

An active and engaged community of instructors and learners, both using and advocating for best practices in effective and reproducible research

- Real database examples
- Access to new tools
- Very detailed introduction

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

- Patient instructors
- Qualified facilitators

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