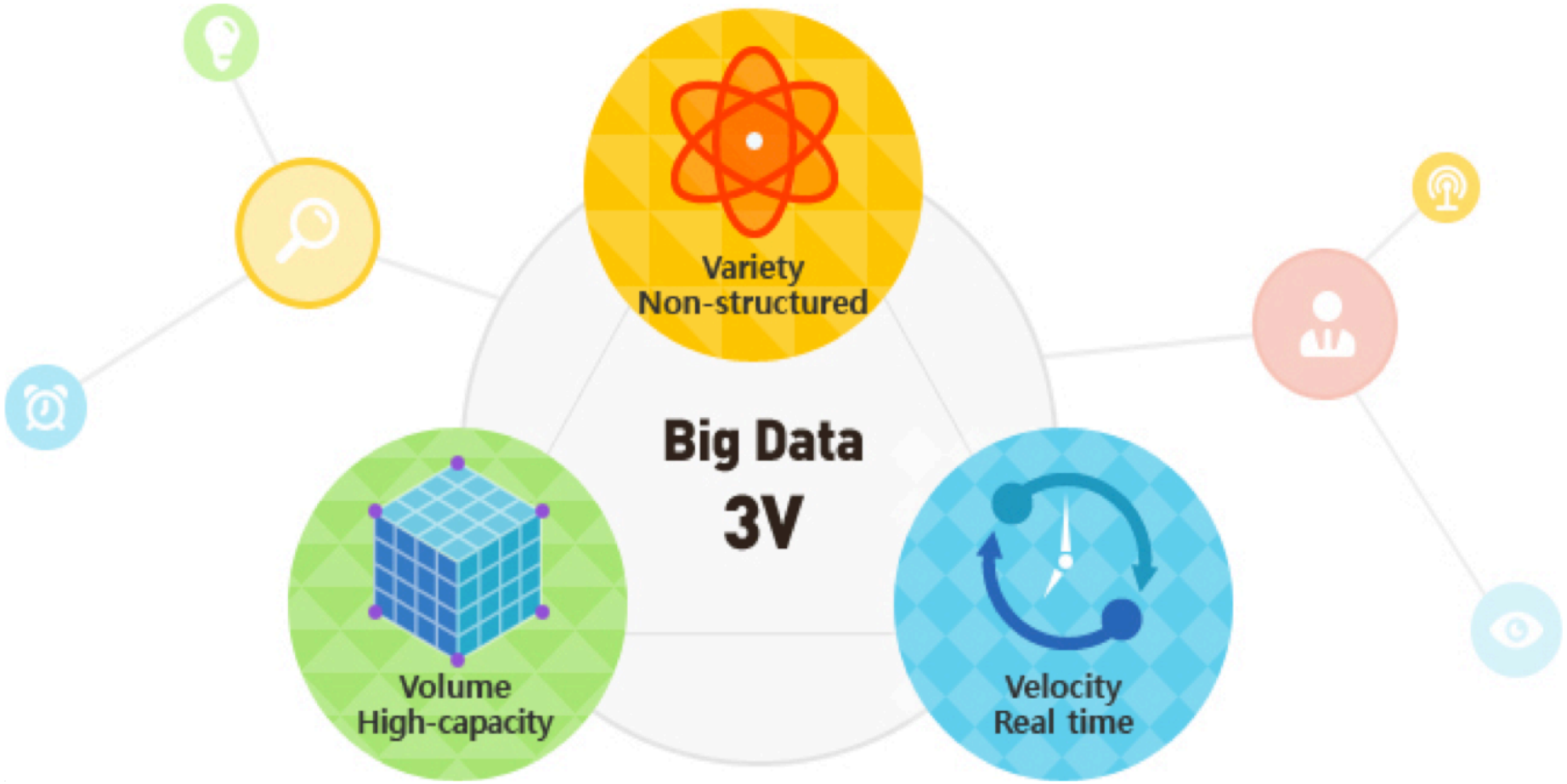
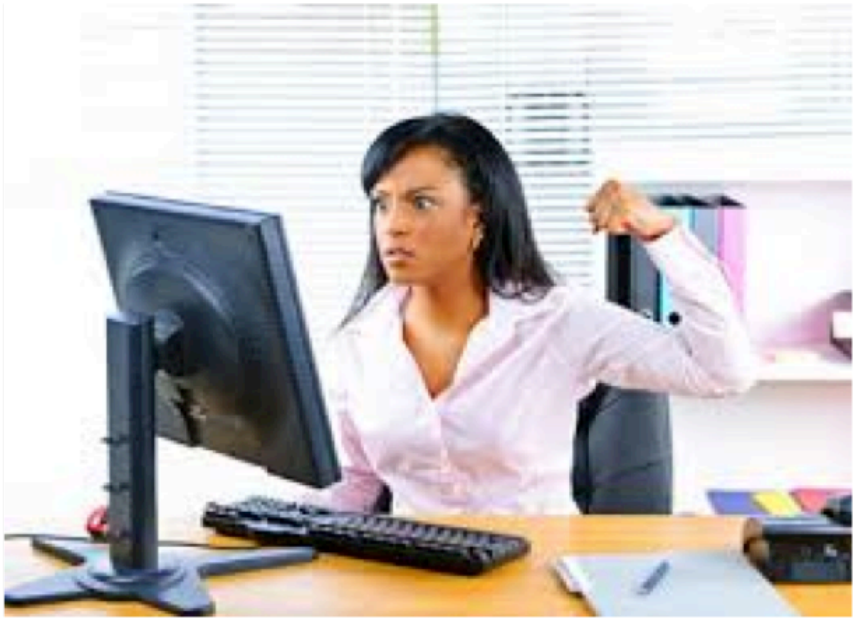
Preventing Bugs in Data Analysis: Data Skills to Improve the Reliability and Effectiveness of Entomological Research

Tracy K. Teal, PhD
Data Carpentry, Executive Director
@tracykteal, @datacarpentry
Find these slides on GitHub: https://github.com/tracykteal/ICE2016-presentation/

This is an exciting time to do research


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



Software and tools allow us to turn data into information.

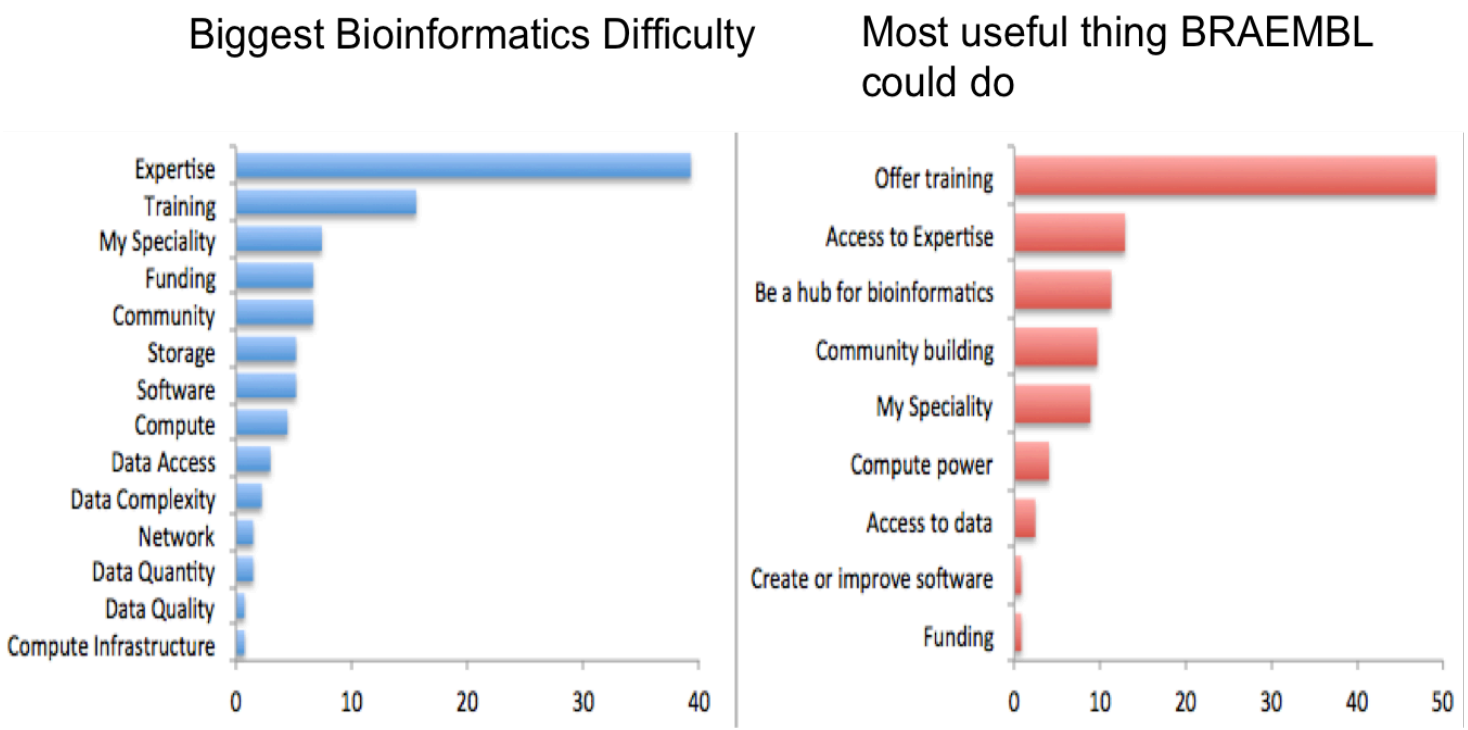
People turn information into knowledge.

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.

Unleash the potential of data by empowering people

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

Researchers want these skills



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

Data Carpentry workshops are for any researcher who has data they want to analyze, and no prior computational experience is required. This **hands-on workshop** teaches basic concepts, skills and tools for working more effectively with data.

- Focused on data - teaches how to manage and analyze data in an effective and reproducible way.
- Initial focus is on training 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 CyVerse and in geospatial data developed with NEON
- Lessons openly licensed and collaboratively developed
- Trained volunteer instructors

What we teach

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.

How we teach: Hands on intensive workshops

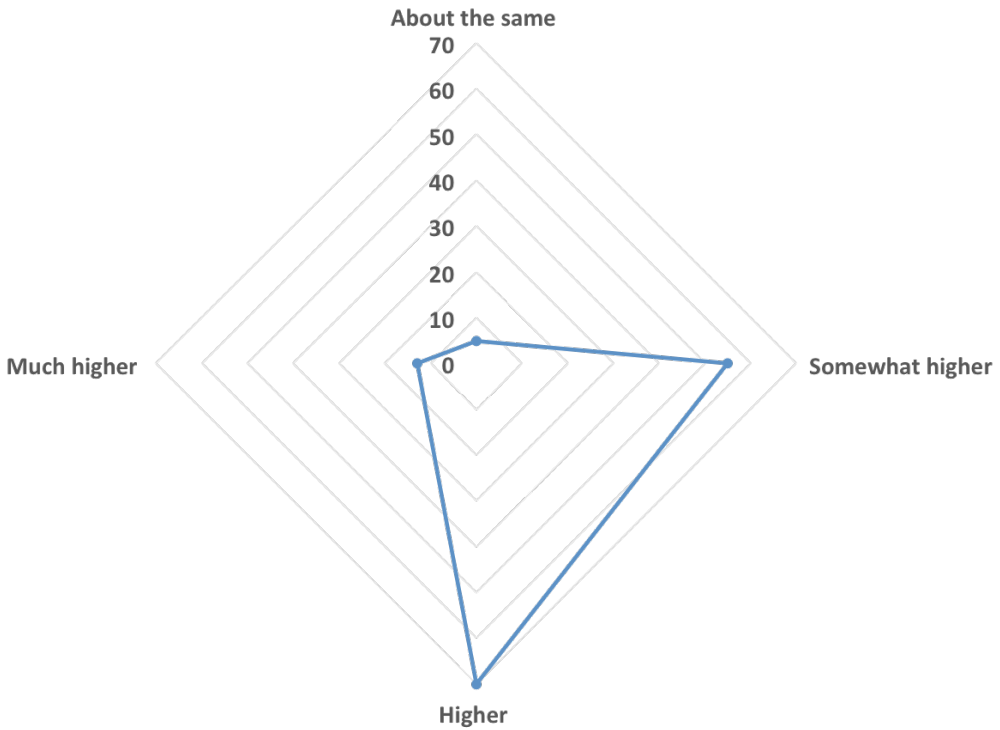
- Short: Two days
- Impactful: Focused time
- Convenient: Held at the university or organization
- Interactive: Hands-on teaching and exercises
- Immediate feedback: sticky notes & minute cards
- Qualified instructors
- Shared learning and a friendly learning environment

Instructors

Outcomes

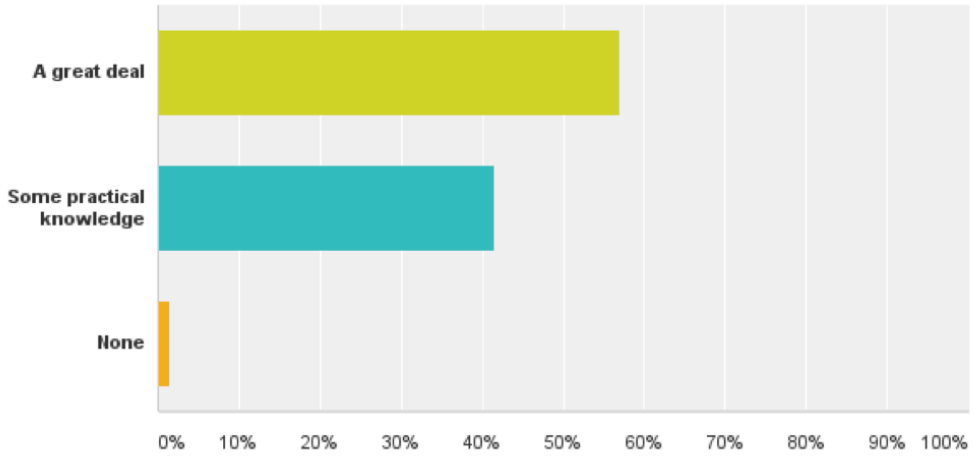
- Learner outcomes
- Instructor outcomes
- Community!

Level of Data Management/Analysis Skills (Post-workshop)



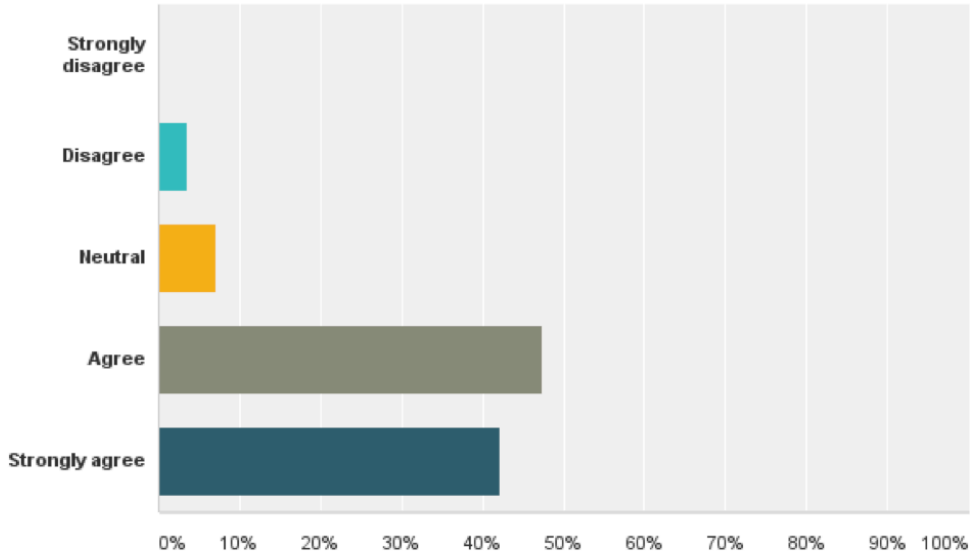
Q8 How much practical knowledge have you gained from this workshop?

Answered: 65 Skipped: 5



Q14 This workshop was worth my time.

Answered: 57 Skipped: 13




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

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

DATA CARPENTRY

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
- 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