Data Carpentry: Data Skills Training to Enable More Effective & Reproducible Research

Tracy K. Teal, PhD

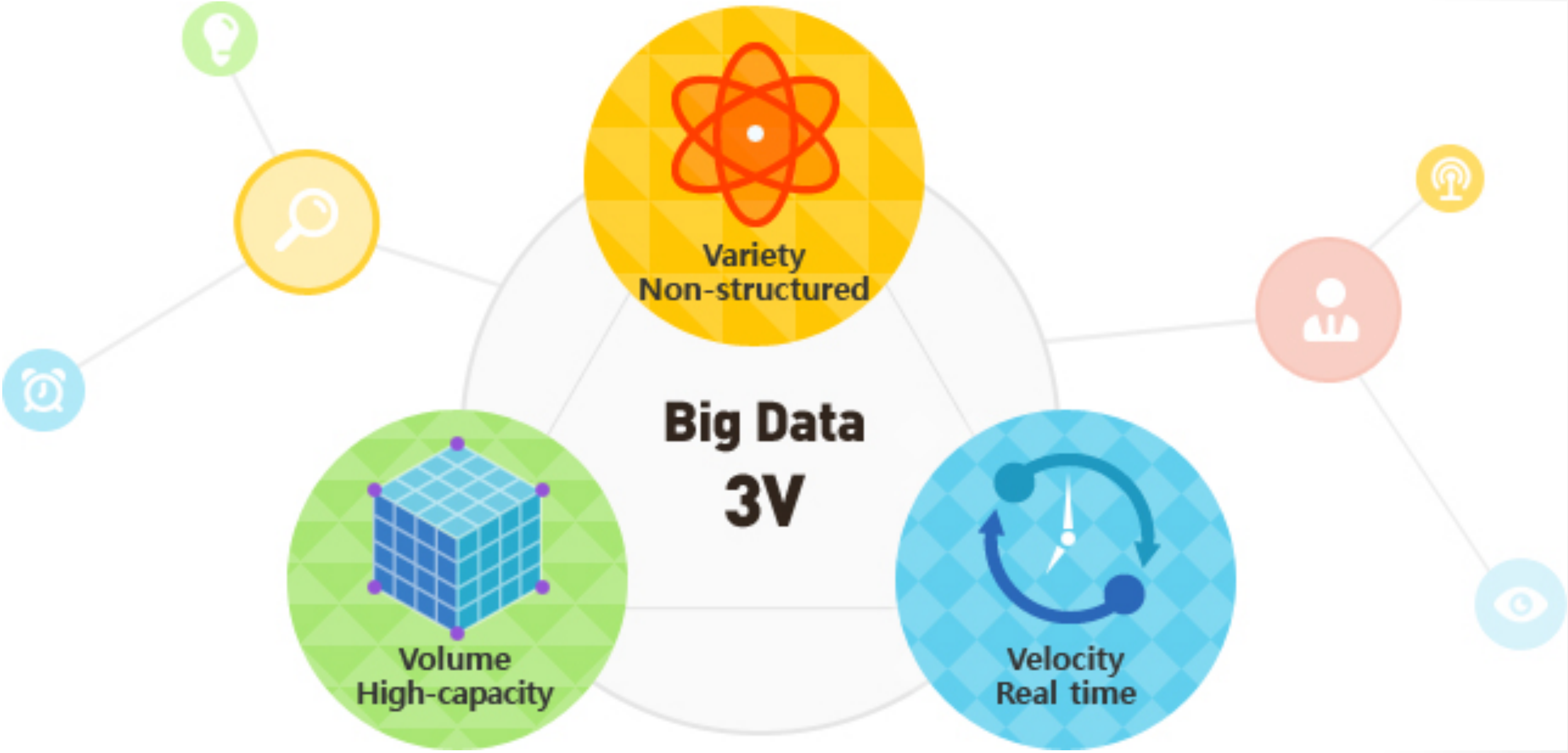
Data Carpentry

Executive Director

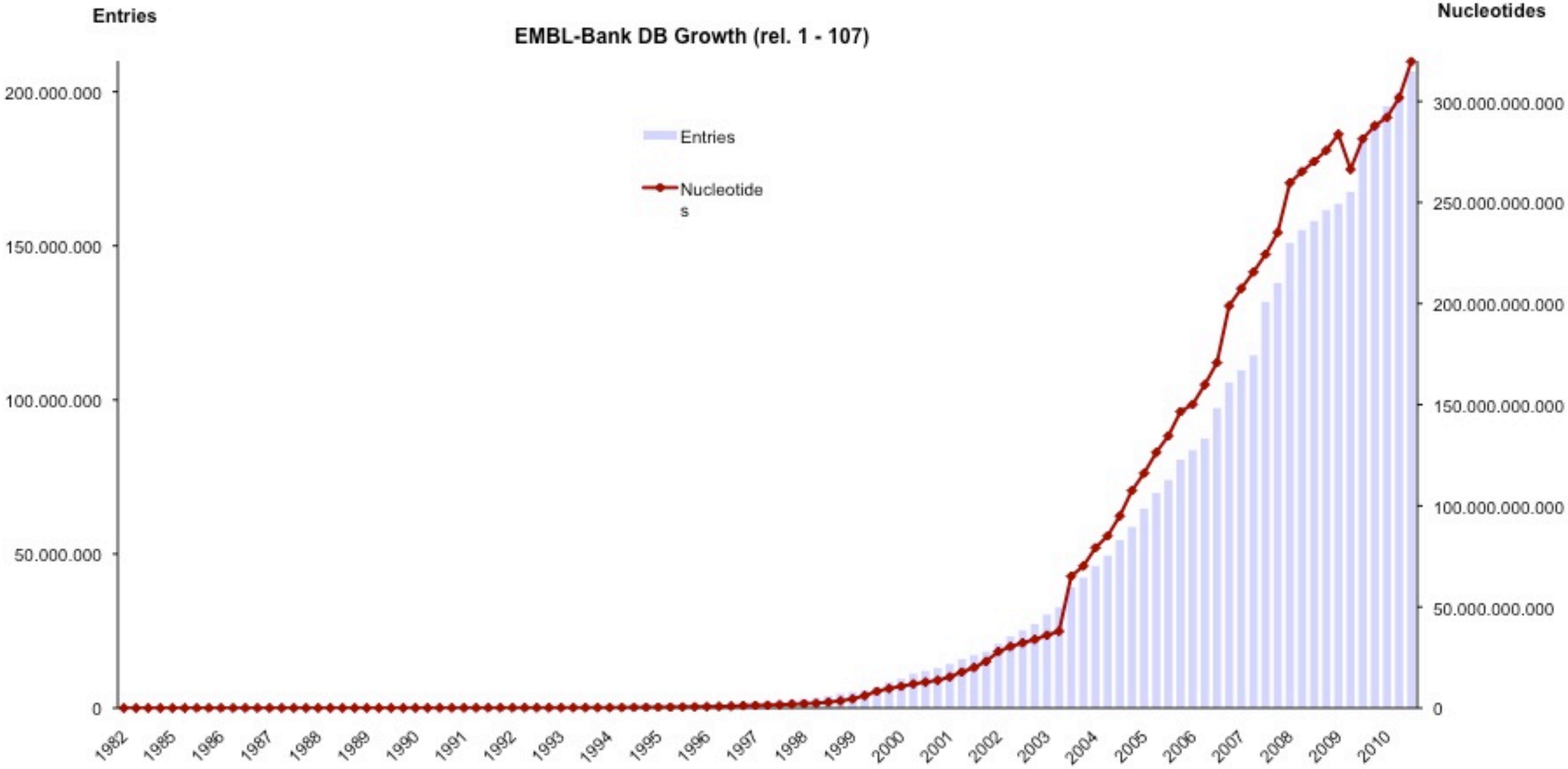
@datacarpentry   @tracykteal

http://www.datacarpentry.org

Our increasing capacity to collect data is changing science



Data production is increasing



Working with 'big data' is no longer the domain of specialists and is instead widely done by all researchers.

People are struggling to manage & analyze this data

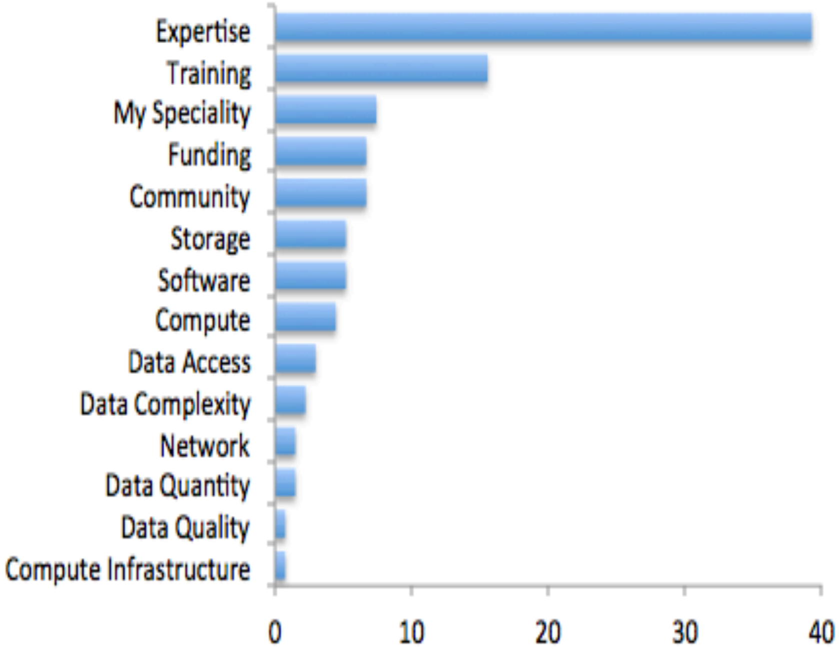
How do we scale ‘data literacy’ along with data production?

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

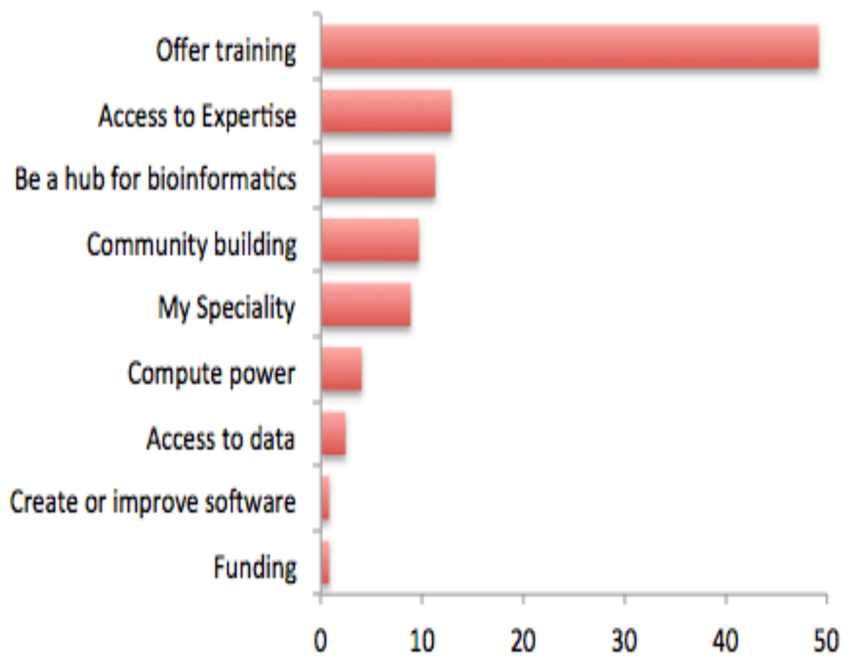
Researchers view the major limiting factor in research progress as a lack of expertise in how to handle and analyze data

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

Where does this training come from?

Data Carpentry workshops to fill that training gap

Our goal is to provide researchers high-quality training covering the fundamentals and best practices in the full lifecycle of data-driven research.

Teach basic concepts, skills and tools for working more effectively with data. Workshops are designed for people with little to no prior computational experience.

DATA CARPENTRY

Founded as a sibling organization in 2014 out of National Science Foundation BIO Centers

Grassroots training effort

- Developed by practitioners for practitioners
- Identify skills and best practices needed in data management and analysis in given domains
- Collaboratively and iteratively developed openly licensed (CC-BY) training materials (all available on github)
- Organize and deliver two-day, intensive hands-on workshops in fundamental data analysis skills using a pool of volunteer helpers and instructors

Hands-on intensive workshops

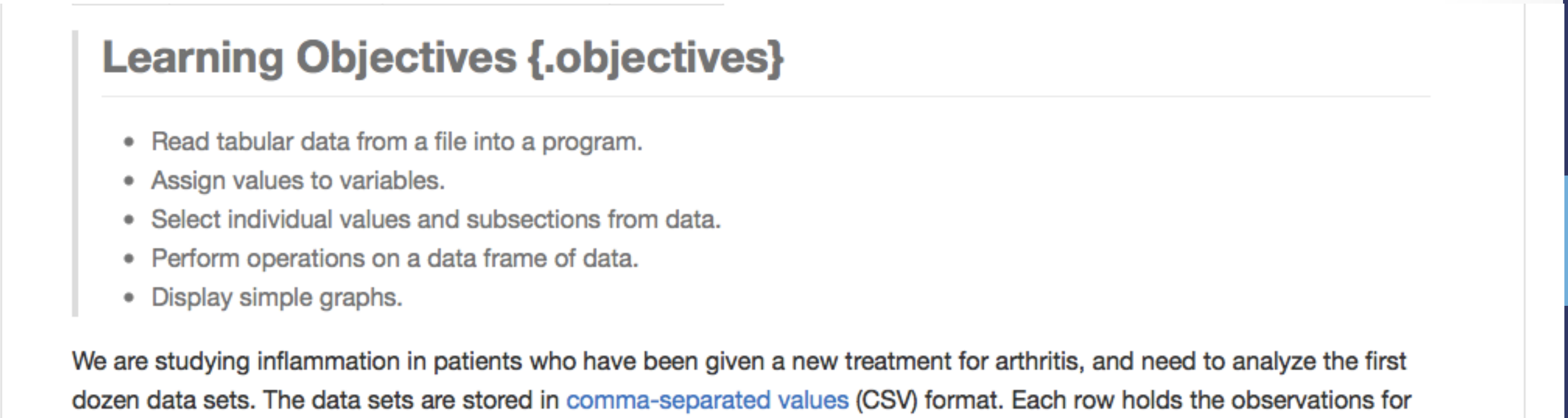
- Two days
- Hands-on
- Qualified instructors
- Helpers
- Post-it notes!
- Friendly learning environment

Image by Aleksandra Pawlik

We can't teach everything in two days, but the goal is to teach foundational skills to reduce the activation energy for getting started and to know what's possible.

Volunteer instructors

Collaboratively developed curriculum



Learning Objectives {.objectives}

- Read tabular data from a file into a program.
- Assign values to variables.
- Select individual values and subsections from data.
- Perform operations on a data frame of data.
- Display simple graphs.

We are studying inflammation in patients who have been given a new treatment for arthritis, and need to analyze the first dozen data sets. The data sets are stored in comma-separated values (CSV) format. Each row holds the observations for

- Focused on data - teaches how to manage and analyze data in an effective and reproducible way.
- Initial focus is on workshops 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 iPlant, and in geospatial data developed with NEON

Curriculum

Domain specific by design,
to focus on relevant data type and skills

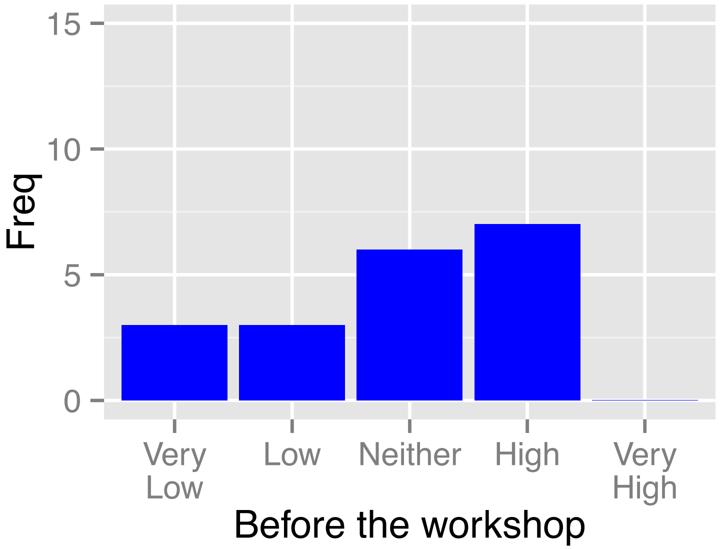
http://www.datacarpentry.org/lessons/

Developing new curriculum

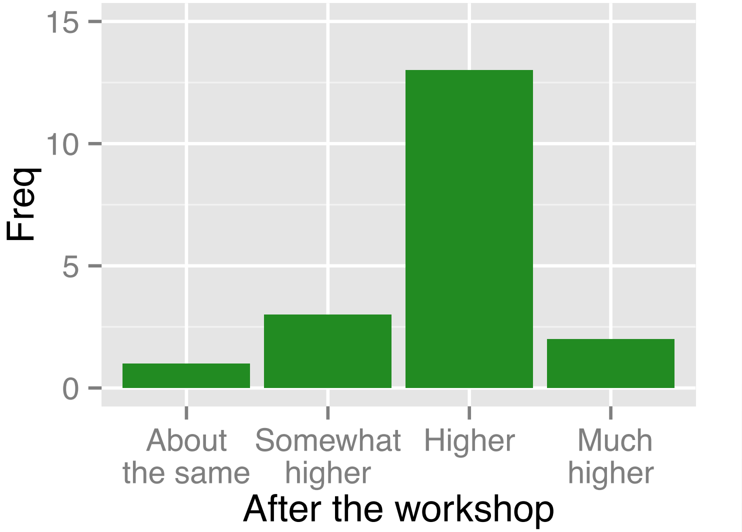
- Initial lessons
- Enthusiasm!
- Hackathons
- Early workshops
- Templating
- Resources for continued support and development

People are learning things!

Level of data management and analysis skills prior to the workshop

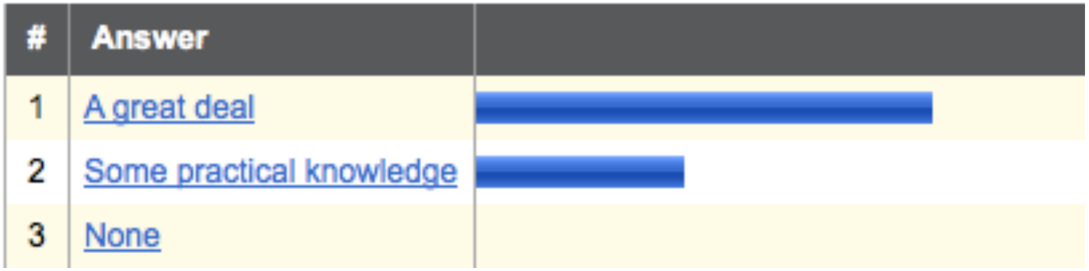


Rate your level of data management and analysis skills following the workshop



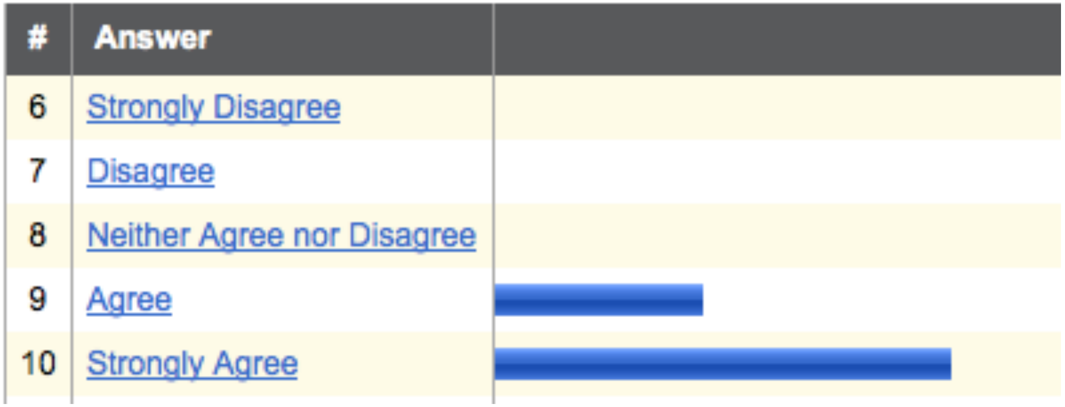
They feel the workshop was worthwhile

How much practical knowledge did you gain from this workshop?



#	Answer	
1	A great deal	
2	Some practical knowledge	
3	None	

This workshop was worth my time



#	Answer	
6	Strongly Disagree	
7	Disagree	
8	Neither Agree nor Disagree	
9	Agree	
10	Strongly Agree	

Get Involved

- Request a workshop
- Be a helper at a workshop
- Become a partner or affiliate and train local instructors
- Contribute to lesson development

Guiding Data Carpentry

Data Carpentry Steering Committee:

Karen Cranston (OpenTree of Life)

Hilmar Lapp (Duke)

Aleksandra Pawlik (Software Sustainability Institute)

Karthik Ram (rOpenSci / Berkeley Institute of Data Science Fellow)

Ethan White (University of Florida / Moore DDD Investigator)

Greg Wilson (Software Carpentry)

Support

Data Carpentry

GORDON AND BETTY MOORE FOUNDATION

BEACON

SESYNC

NESCent National Evolutionary Synthesis Center

DataONE Data Observation Network for Earth

iDigBio Integrated Digitized Biocollections

iPlant Collaborative™

Software Carpentry

http://software-carpentry.org/scf/partners/

Jason Williams at iPlant