iDigBio
Integrated Digitized Biocollections

DATA CARPENTRY

Got data? Need data skills? Check out datacarpentry.org

Presenter: Deborah Paul, @idbdeb
Florida State University
Integrated Digitized Biocollections (iDigBio)

at Entomological Collections Network 2014
Portland, Oregon 15 November 2014
Author: Deborah Paul

#datacarpentry
#EntColl2014

UF UNIVERSITY of FLORIDA
FLORIDA STATE UNIVERSITY 1851
FLORIDA MUSEUM OF NATURAL HISTORY
NSF

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Researchers are experiencing a lot of data pain and are frustrated or limited by their current workflows

DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Goal:

Develop and teach workshops to help *train current and next generation researchers good data analysis and management practices to enable individual research progress and open and reproducible research*.

What is Data Carpentry?

Two-day (Three-day?) intensive workshops, modeled on Software Carpentry

Learning objective:

Researchers should be able to retrieve, view, manipulate, analyze and store their and other's data in an open and reproducible way.

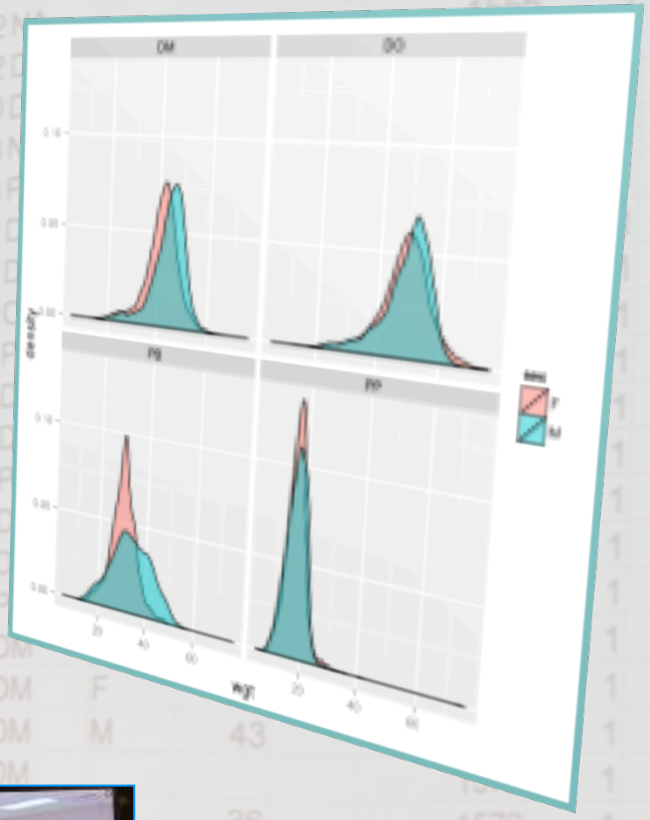
- Data Carpentry is focused on data - The workshop introduces one data set at the beginning of the workshop. This data set is used throughout the workshop to teach how to manage and analyze data in an effective and reproducible way.
- Data Carpentry is designed for novices - there are no prerequisites, and no prior knowledge about the tools is assumed.
- Data Carpentry is domain specific by design.

Sentiments on data within the NSF BIO Centers (BEACON, SESYNC, NESCent, iPlant, iDigBio)

DATA CARPENTRY

- I usually manage data in Excel and it's terrible and I want to do it better.
- I'm organizing GIS data and it's becoming a nightmare.
- My advisor insists that we store 50,000 barcodes in a spreadsheet, and something must be done about that.
- I'm having a hard time analyzing microarray, SNP or multivariate data with Excel and Access.
- I want to use public data.
- I work with faculty at undergrad institutions and want to teach data practices, but I need to learn it myself first.
- I'm interested in going in to industry and companies are asking for data analysis experience.
- I'm trying to reboot my lab's workflow to manage data and analysis in a more sustainable way.
- I'm re-entering data over and over again by hand and know there's a better way.
- I have overwhelming amounts of data.
- I'm tired of feeling out of my depth on computation and want to increase my confidence.

Scenes from a Data Carpentry Workshop at iDigBio and the AMNH

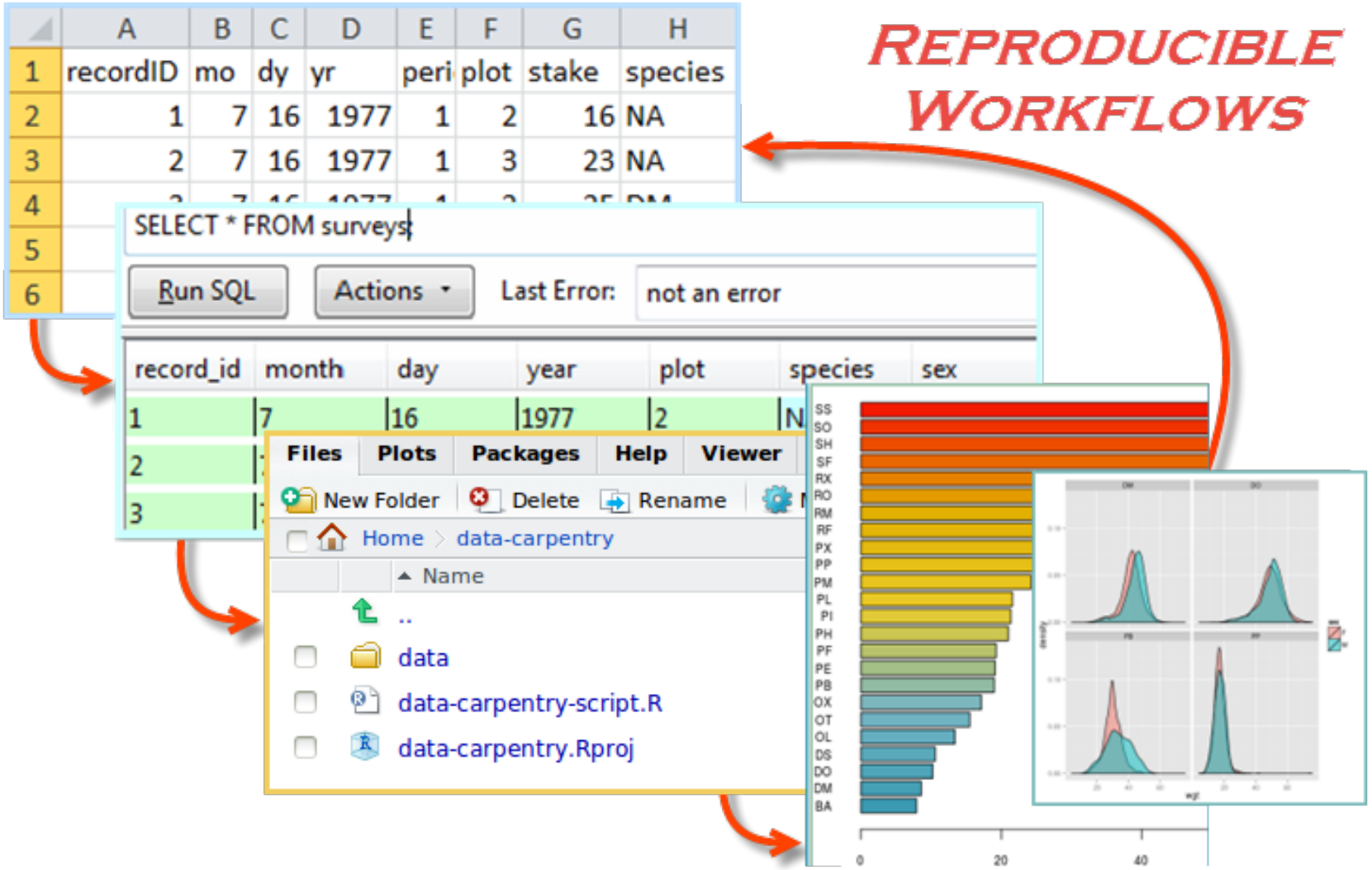


A Typical Data Carpentry Workshop

Course Overview - Day 1		
8:30-8:45	Introductions & Overview, Data Carpentry: Making data science more efficient	All, Deb Paul
8:45-9:00	Linking Heterogeneous Data in Biodiversity Studies: the need for data carpentry	Pam Soltis, iDigBio PI
9:00-10:00	Better use of spreadsheets, part I	Tracy Teal
10:00-10:30	Break	
10:30-12:00	Better use of spreadsheets part II	Tracy Teal
12:00-1:30	Lunch (with OpenRefine Demo)	Deb Paul
1:30-3:00	SQL Introduction	Matt Collins
3:00-3:30	Break	
3:30-5:00	SQL part II	Matt Collins
5:00-5:30	Review / Wrap up for tomorrow	
Course Overview - Day 2		
8:30-10:00	Introduction to the shell	Tracy Teal
10:00-10:30	Break	
10:30-12:00	Introduction to R	François Michonneau
12:00-1:30	Lunch	
1:30-3:00	Manipulating and plotting data in R	François Michonneau
3:00-3:30	Break	
3:30-4:30	Getting data in and out of R: How to integrate R in your workflow	François Michonneau
4:30-5:00	Sharing your data and your results: RMarkdown and Figshare	François Michonneau
5:00-5:30	Review / Wrap up / Evaluation and Feedback	

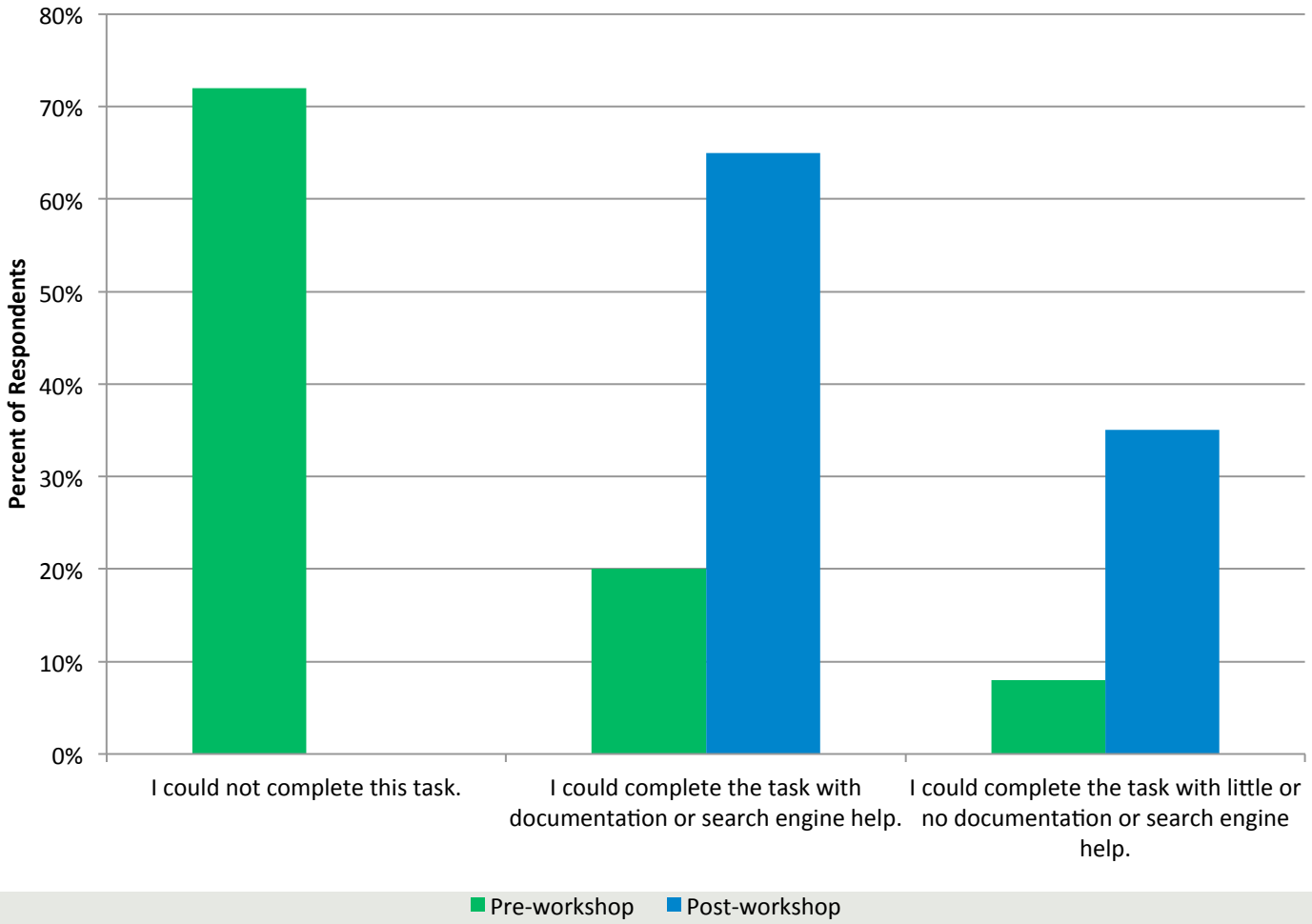
https://www.idigbio.org/wiki/index.php/Data_Carpentry

Reproducible Workflows



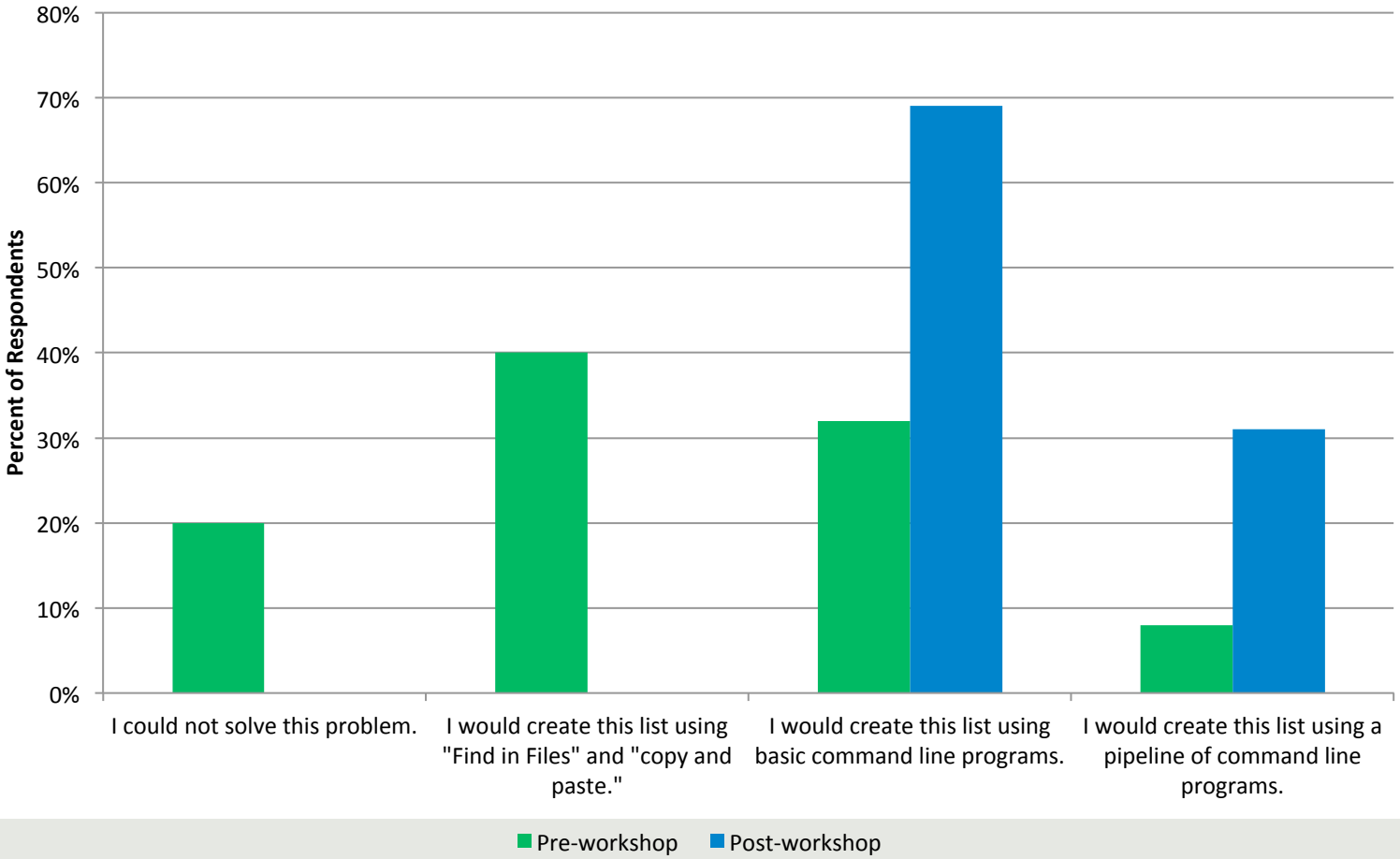
Assessment

Consider this task: A database has two tables: Scientist and Lab. Scientist's columns are the scientist's user ID, name, and email address; Lab's columns are lab IDs, lab names, and scientist IDs. Write an SQL statement that outputs the number of scientists in each lab.



Assessment

How would you solve this problem? A directory contains 1000 text files. Create a list of all files that contain the word “Drosophila” and save the result to a *file* called results.txt.



Data Carpentry curriculum

- Preparing data for analysis
- How to organize data and use spreadsheet programs more effectively, but also to recognize their limitations.
- Getting data out of spreadsheets and into tools such as R or Python that allow for reproducible workflows and have more capabilities.
- Using databases, including managing and querying data in SQL.
- Workflows and automating repetitive tasks, in particular using the command line shell and shell scripts.
- Using data and computational resources, in particular publicly available ones such as Amazon, DataDryad and Figshare
- Overall, conducting data and computation-heavy research more efficiently, reproducibly and openly.

Data Carpentry instructor development and resources

- Training and supporting instructors is another primary goal of Data Carpentry
- Providing open source/creative commons materials for re-use
- Potentially acting as a hub for instructional materials on data analysis and management

Materials development

Currently materials for multiple domains and topics and working with people in different domain to develop more

Topics:

Shell, R, Python, SQL, Excel, data cleaning, text mining, HDF5

Domains:

Ecology, genomics, social science, neuroscience, geosciences

Community driven effort

Data Carpentry board:

Karen Cranston (NESCent), Hilmar Lapp (Duke), Aleksandra Pawlik (ELIXIR UK), Karthik Ram (rOpenSci), Tracy Teal (Michigan State), Ethan White (Univ of Florida), Greg Wilson (Software Carpentry)

Contributors:

20 people contributing to materials development already

4 workshops taught, 11 instructors, ~20 helpers

Open source materials

https://github.com/datacarpentry/datacarpentry/

Next bits for Data Carpentry

- Field-to-Database Workshop at iDigBio
- Hackathon (API development)
- Assessment follow up
- Join us, Request a workshop, become an instructor,…
- Share your ideas, materials, models, courses
- What about other initiatives?
  - Biodiversity Information Standards (TDWG) Interest Group
- Talk to us!

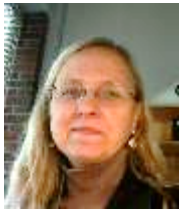
DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Thank you ECN 2014!

- Tracy K Teal, Michigan State University
- Francois Michonneau, iDigBio Post Doc
- Katja Seltmann, AMNH, TTD - TCN
- Matt Collins, iDigBio
- Kevin Love, iDigBio
- Reed Beaman, iDigBio
- SESYNC, iPlant, BEACON, NESCent,
- And the Data Carpentry Board

Work presented here made possible by many and especially…



dpaul@fsu.edu
@idbdeb

Find out more at http://www.datacarpentry.org

www.idigbio.org

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data Literacy and Computational Literacy