Data Carpentry: Enabling Researchers to Work More Effectively with Data

@datacarpentry
@thecarpentries
http://www.datacarpentry.org

- General overview
- Introductions of instructors and attendees
- Logistics
- Data Carpentry Genomics!

With the emergence of new technologies generating large datasets in all domains of research, data analysis is no longer the domain of specialists and is instead widely done by all researchers.

Bioinfo training gap

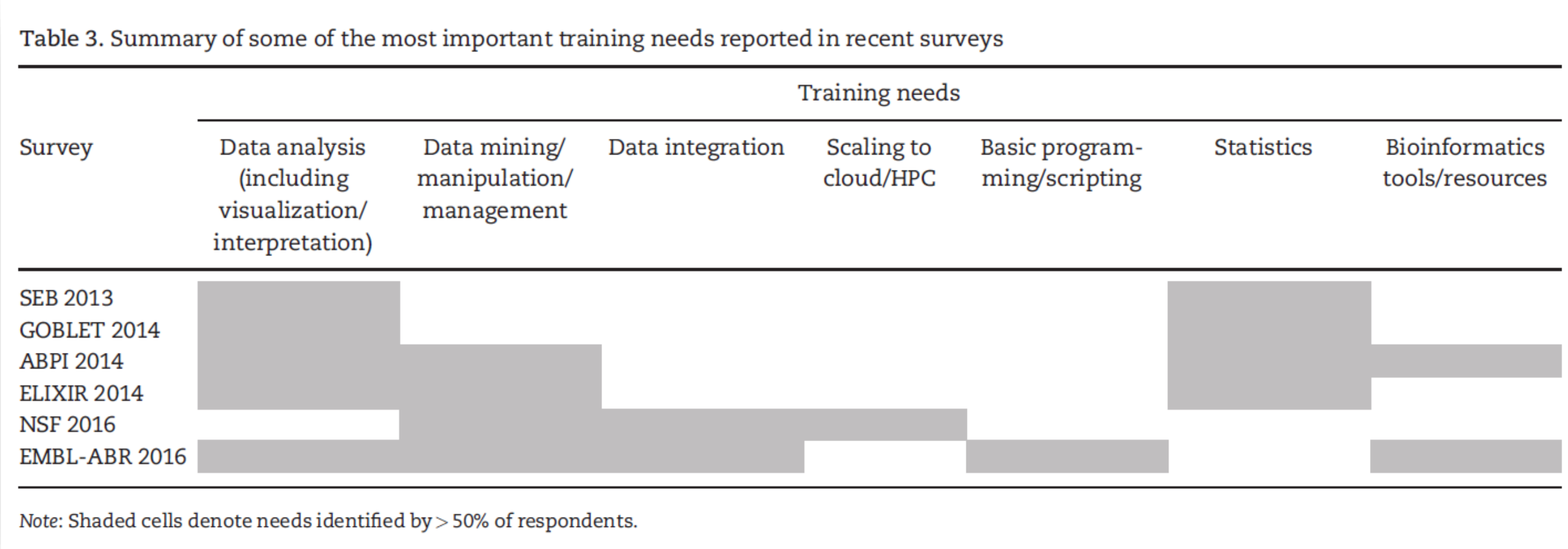

Table 3. Summary of some of the most important training needs reported in recent surveys

Survey	Training needs						
	Data analysis (including visualization/ interpretation)	Data mining/ manipulation/ management	Data integration	Scaling to cloud/HPC	Basic programming/scripting	Statistics	Bioinformatics tools/resources
SEB 2013	▨					▨	
GOBLET 2014	▨					▨	
ABPI 2014	▨	▨				▨	▨
ELIXIR 2014	▨	▨				▨	
NSF 2016		▨	▨	▨			
EMBL-ABR 2016	▨	▨	▨		▨		▨

Note: Shaded cells denote needs identified by > 50% of respondents.

Attwood, T.K. et al. *Briefings in Bioinformatics,* 2017, 1-7

Sentiments on data within the NSF BIO Center

- I'm having a hard time analyzing microarray, SNP or multivariate data with Excel and Access.
- I want to use public data.
- 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.

Data Carpentry is trying to help fill that training gap

Our goal is to provide researchers training on the fundamentals and best practices in data analysis and management.

(i.e. the stuff that doesn't get taught in most classes)

DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Data Carpentry workshops

We know 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 for you to know what’s possible.

About this workshop

- No prerequisites, and no prior computational experience is assumed.

- A friendly environment for learning!

  We have a code of conduct, but basically be respectful of each other and instructors.

- Interactive learning!

  Lessons will all be hands on and let you work through the materials. Please ask questions! Talk to your neighbor, ask a helper, the instructor or put notes in the etherpad.

Hello!

Instructors

- **Catalina Anghel** (Canadian Nuclear Labs)
- **Ahmed Hasan** (UTM Dept. of Cell and Systems Biology)
- **Heather Gibling** (Ontario Institute for Cancer Research and UofT Dept. of Molecular Genetics)

Helpers

Centre for Addiction and Mental Health

- **Erin Dickie**
- **Marcia Hon**
- **Ricardo Harripaul**
- **Nikola Bogetic**
- **Arin Bakht**

Before we start

- **Website:** https://canghel.github.io/2018-05-22-camh/
  - Links to lessons

- **Etherpad:** http://pad.software-carpentry.org/2018-05-22-camh
  - Online discussion forum
  - Add links, get help with questions, copy code into window

Sticky notes

Other things…

- Bathrooms
- Breaks
- Can see screen?

Acknowledgements

Slides adapted from:

- Jason J Williams: https://github.com/JasonJWilliamsNY/2017-08-28-nmsu/blob/gh-pages/DataCarpentry_Intro.pdf
- Erica Nudrak, https://github.com/emudrak/2017-06-14-cornell/tree/gh-pages/slides

Images

- http://acobiom.com/bioinformatics/