

Data Carpentry: Enabling Researchers to Work More Effectively with Data

Tracy K. Teal, PhD

Data Carpentry Project Lead

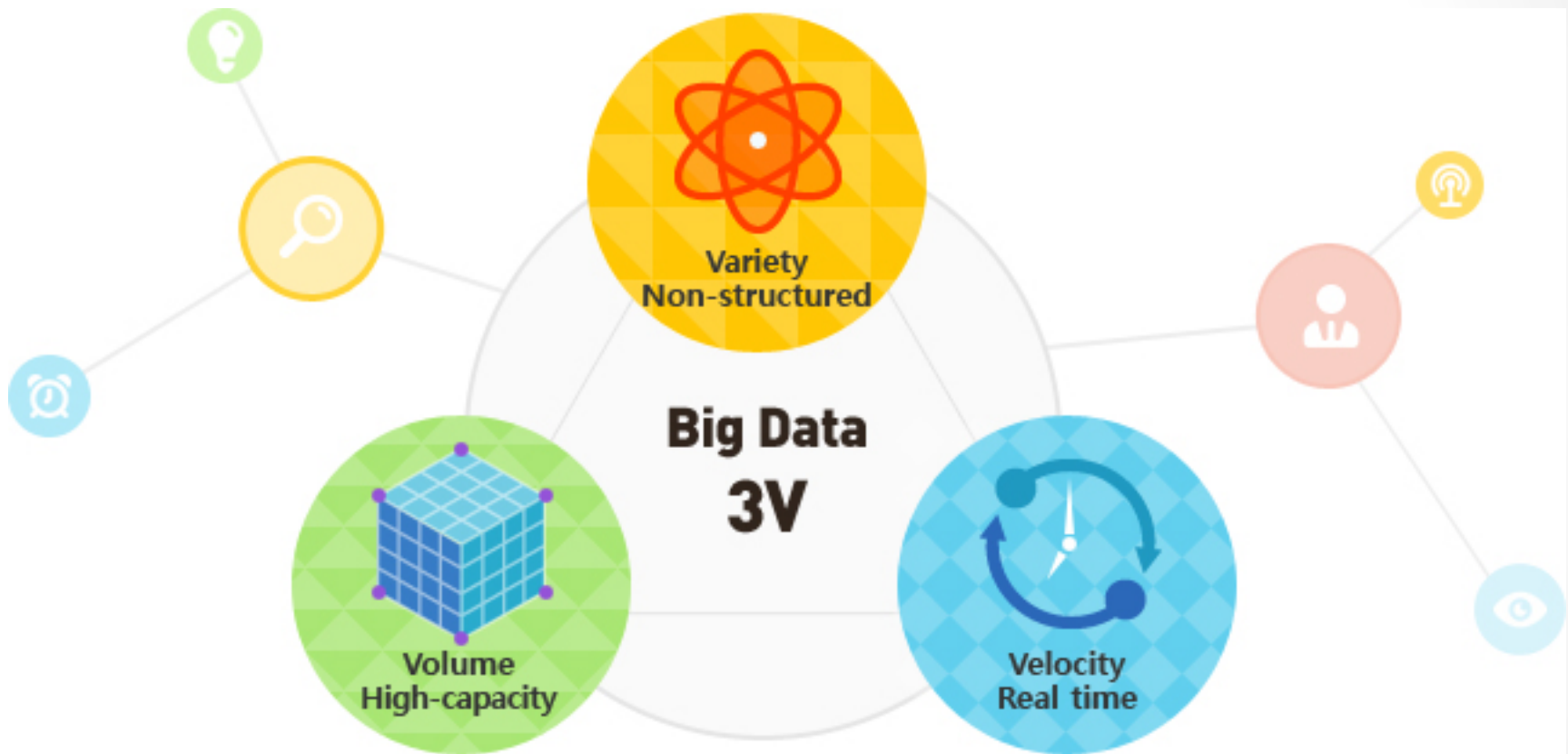
Assistant Professor, BEACON, Michigan State University

@datacarpentry

<http://datacarpentry.org>

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





Large scale data is being generated in all domains



As well as in the non-academic sector



Data potential

Data is not
information

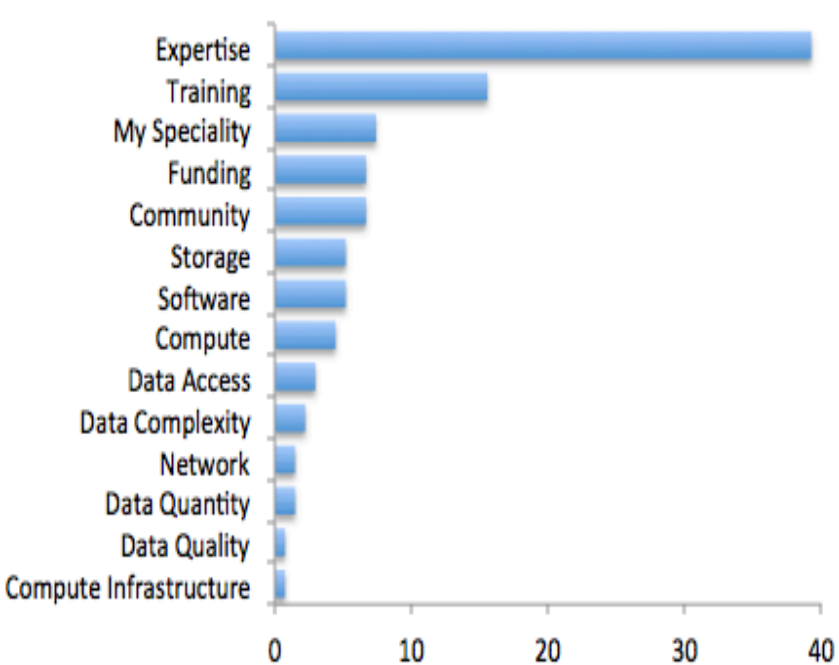
Training is a missing piece between
data collection & data-driven discovery



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



Data Carpentry is filling that training gap

Our mission is to provide researchers high-quality,
domain-specific training covering the full lifecycle
of data-driven research.



DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

We're here to help

(the logo is a saw)



- Training focused on data - teaching how to manage and analyze data in an effective and reproducible way.
- Domain specific by design – currently have lessons in ecology and are developing lessons for genomics, geosciences and social sciences.
- Initial focus is on novices - there are no prerequisites, and no prior knowledge computational experience is assumed. We plan to expand to more advanced topics.

Grassroots training effort

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

Grassroots training effort

- Developed by practitioners for practitioners
- Identify skill needs in data management and analysis in given domains
- Collaboratively and iteratively developed openly licensed (CC-BY) [training materials](#)
- Organize and deliver two-day, intensive hands-on [workshops](#) in fundamental data analysis skills using a pool of volunteer helpers and instructors

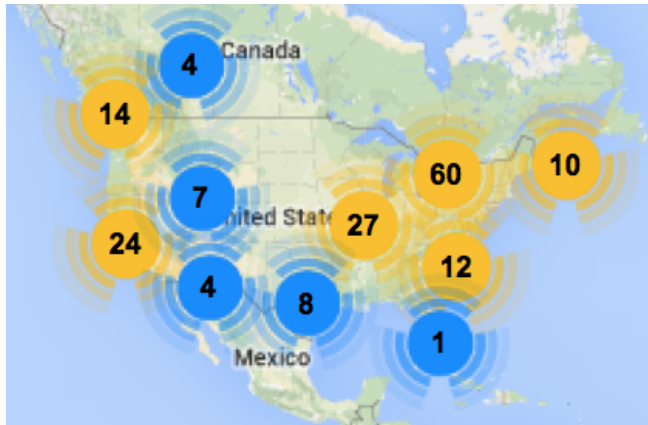
Grassroots training effort

- Developed by practitioners for practitioners
- Identify skill needs in data management and analysis in given domains
- Collaboratively and iteratively developed openly licensed (CC-BY) [training materials](#)
- Organize and deliver two-day, intensive hands-on [workshops](#) in fundamental data analysis skills using a pool of [volunteer helpers and instructors](#)

Software Carpentry



Greg Wilson founded in 1998



*North American Workshops
2012-2014*

Since January 2013

With the help of the
Mozilla Science Lab
scaled to teach

- Over 270 two-day workshops
- For over 8300 learners
- Taught by over 200 volunteers
- In over 20 countries

Now its own non-profit the Software Carpentry Foundation

Data Carpentry workshops

Goals:

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 know what's possible

Curriculum: The data lifecycle from data organization to analysis and visualization

Data Carpentry workshops

Format

- Two days
- Hands on
- Qualified instructors
- Helpers
- Sticky notes!



Demand is high

Workshops internationally

Started in November, 2014; since Jan 2015 have taught 10 workshops and have more than 24 scheduled for this year

Interest from broad domains – biology, genomics, social science, digital humanities, libraries, geosciences

Curriculum for ecology

The data lifecycle from data organization to analysis and visualization

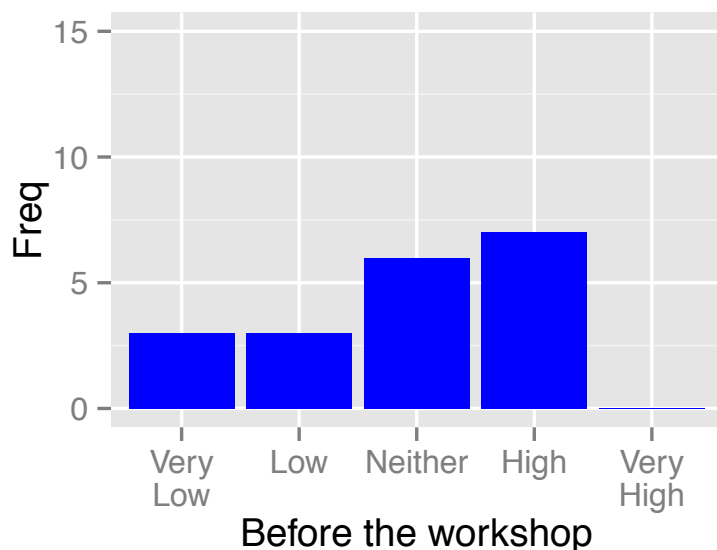
- Data organization in spreadsheets
- OpenRefine for data cleaning
- R for data analysis and visualization
- SQL

Workshop at NDIC

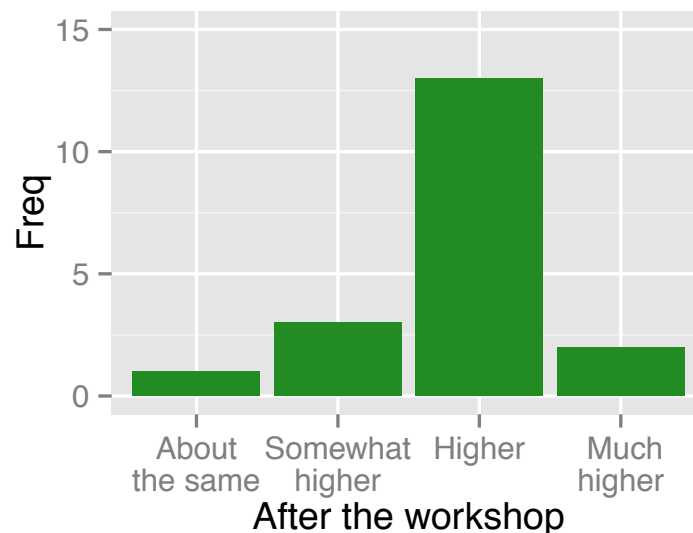
<https://github.com/datacarpentry/2015-05-03-NDIC/wiki>

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





People are learning things!

Compared to before the workshop, I have a better understanding of how to



#	Question	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1	Effectively organize data in spreadsheets	<u>0</u>	<u>1</u>	<u>7</u>	<u>4</u>	<u>7</u>
2	Use OpenRefine for data cleaning	<u>0</u>	<u>0</u>	<u>0</u>	<u>9</u>	<u>10</u>
3	Import a data file in to R & work with the data	<u>0</u>	<u>0</u>	<u>2</u>	<u>7</u>	<u>10</u>
4	Do initial visualizations in R	<u>0</u>	<u>0</u>	<u>3</u>	<u>6</u>	<u>10</u>
5	Construct an SQL query statement	<u>0</u>	<u>0</u>	<u>3</u>	<u>7</u>	<u>9</u>

They feel the workshop was worthwhile

How much practical knowledge did you gain from this workshop?

#	Answer		Response
1	A great deal		13
2	Some practical knowledge		6
3	None		0

This workshop was worth my time

#	Answer		Response
6	Strongly Disagree		0
7	Disagree		0
8	Neither Agree nor Disagree		0
9	Agree		6
10	Strongly Agree		13

Thoughts on data best practices

Please rate your level of agreement with the following statements

#	Question	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
2	Data organization is a fundamental component of effective and reproducible research.	<u>0</u>	<u>0</u>	<u>0</u>	<u>1</u>	<u>18</u>
3	Using a scripting language like R or Python can ultimately improve my analysis efficiency	<u>0</u>	<u>0</u>	<u>1</u>	<u>6</u>	<u>12</u>
4	Using R or Python makes analyses easier to reproduce	<u>0</u>	<u>0</u>	<u>3</u>	<u>5</u>	<u>11</u>
6	A value of using SQL, R or Python is that the underlying data cannot accidentally be changed	<u>0</u>	<u>1</u>	<u>1</u>	<u>7</u>	<u>10</u>

Hackathons to develop lessons

- Genomics

project organization, command line, cloud computing, using bioinformatics tools, data analysis and visualization

- CSHL, iPlant, SESYNC, iDigBio

- Geospatial data

Working with geospatial data

Hackathon at NEON – Sept/Oct (Leah Wasser)

- Social sciences

Working with data from social sciences

Hackathon at Berkeley – July (Dav Clark)

What Students Learn

- Capstone lesson

Guiding Data Carpentry

Steering Committee:

Karen Cranston (NESCent / OpenTree of Life)

Hilmar Lapp (NESCent / Duke)

Aleksandra Pawlik (Software Sustainability Institute)

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

Tracy Teal (Data Carpentry / Michigan State)

Ethan White (University of Florida / Moore DDD Investigator)

Greg Wilson (Software Carpentry)

Volunteer instructors and materials developers

Mike Smorul (SESYNC), Mary Shelly (SESYNC), Jason Williams (iPlant), Leah Wasser (NEON), Deb Paul (iDigBio), Francois (iDigBio), Ben Marwick (University of Washington), Dav Clark (Berkeley)

Data Carpentry support



GORDON AND BETTY
MOORE
FOUNDATION



NATIONAL
SOCIO-ENVIRONMENTAL
SYNTHESIS CENTER

