Data Instruction: Developing New Roles for Data Librarians

Tim Dennis
Data Services Librarian

Reid Otsuji
Data Curation Specialist

AAAS-PD Library Science Symposium
June 14, 2016

The Library
UC SAN DIEGO

"Computing is now an integral part of every aspect of science, but most scientists are never taught how to build, use, validate, and share software well." -- Software Carpentry <http://software-carpentry.org/>

Overview

Background

- New data sources
- Emerging research methods requiring data science
- Data related academic majors/programs

How we are responding at UC San Diego

- Offering data workshops
- Working with Software Carpentry
- Engaging with researchers and data communities

The Library
UC SAN DIEGO

Data Science in the Curriculum

UC San Diego
Jacobs School of Engineering

COGS 9: Introduction to Data Science

tics, Spring 2016

Data Science Student Society

ence in Business Analytics (MSBA)

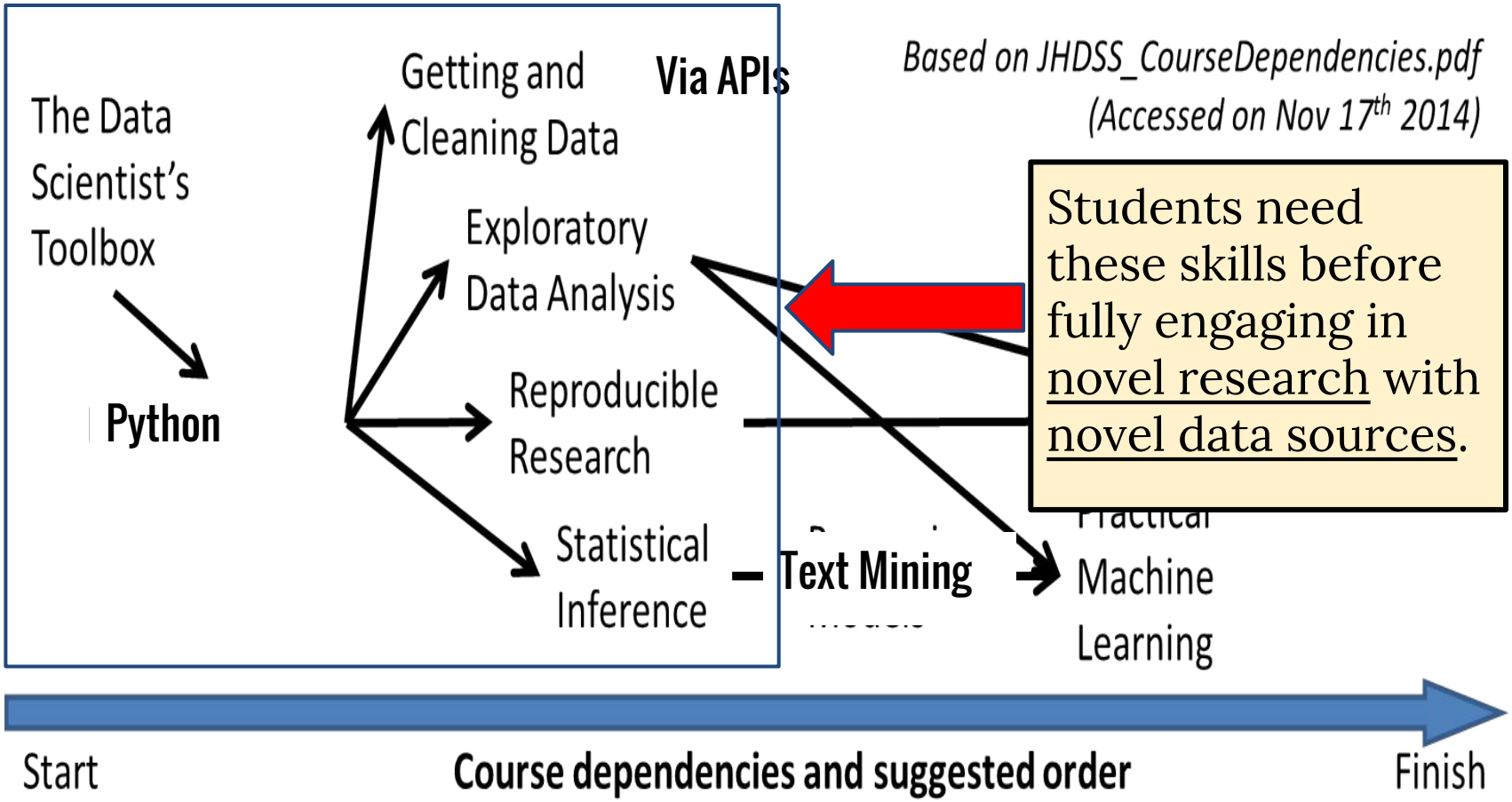
The Library
UC SAN DIEGO

Novel Research Environment

The Library UC SAN DIEGO

Data Sources	**Techniques**
Sensor Data	Big Data Techniques
Social Media	Data Mining
Public Web	Text Mining
Machine Log	Social Network Analysis
APIs	Machine Learning
Text Corpora	Reproducible Research

New Research Depends on Basic CS Skills



Based on JHDSS_CourseDependencies.pdf (Accessed on Nov 17th 2014)

Students need these skills before fully engaging in novel research with novel data sources.

The Library UC SAN DIEGO

Source: https://goo.gl/t9DiBW

Data Workshops

The Library
UC SAN DIEGO

Computational Basics

Identified researchers need for core computing skills.

Focused on providing high-impact novice lessons for the following topics:

- Coding, software development and data management best practices
- Exploratory data analysis
- Working with data (data cleaning)

The Library
UC SAN DIEGO

Software/Data Carpentry

- Goals are to teach core research computing skills
- Lesson materials are freely reusable under [CC BY 4.0]
- Used Software Carpentry/Data Carpentry as a primary resource for lessons and instruction development

Lesson	Site	Repository
The Unix Shell		
Version Control with Git		
Version Control with Mercurial		
Using Databases and SQL		
Programming with Python		
Programming with R		
R for Reproducible Scientific Analysis		
Programming with MATLAB		
Automation and Make		
Instructor Training		

http://software-carpentry.org/lessons/

Instructors

Library program collaboration between:

Library Data Services - Data Librarian

Library Research Data Curation Program - Data Curation Specialist Librarian & Metadata Specialist

Library Information Technology Services - Manager of Development/Web Services

Campus partners (local software carpentry instructors)

Research Data Curation Program

RRAS Data Services

The Library UC SAN DIEGO

Technologies We Chose

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Anaconda

GitHub

python™

>BASH

RStudio

R

jupyter

The Library
UC SAN DIEGO

Data Workshop Instruction Goals

Our goal is to expand the library's research data support and education.

Target audiences:

- Graduate Students
- Researchers
- Research Staff

Types of workshops:

- One shot
- Boot Camp Style

Discipline focused:

- Bioinformatics/Genomics
- Economics
- Global Policy and Strategy
- Academic departments and research labs

The Library
UC SAN DIEGO

One Shot Workshops and Boot Camps

http://ucsdlib.github.io/workshops/

DATA
WORKSHOPS

about

posts

Library Carpentry	July 18, 2016
Jupyter Jumpstart – An Introduction to Literate Programming	June 6, 2016
Python for IASSIST	May 31, 2016
Software Carpentry Workshop	May 17, 2016
Workshop for increasing openness and reproducibility in quantitative research	April 29, 2016
Python for the Economics Department	April 20, 2016
Data Visualization in R for Genomics	March 9, 2016
Data Manipulation in R for Genomics	March 8, 2016
Intro to R for Genomics	March 7, 2016
Data Management 101	February 17, 2016
Intro to R for Genomics	February 9, 2016
Python Programming	January 25, 2016
Intro to R	January 4, 2016
Introduction to Apache Spark and PySpark	November 16, 2015
Data Manipulation, Analysis and Visualization in R	November 15, 2015
Intro to R	November 3, 2015
Software Carpentry Workshop	October 27, 2015

Disciplined Focused Data Instruction

School of Global Policy and Strategy:

- Winter 2016, 9 week skills course series in R, Python, SQL, and data management 101
- Total enrollment: 122 GPS masters students
- Gave out 93 certificates of proficiency

http://ucsdlib.github.io/win2016-gps-intro-R/

http://ucsdlib.github.io/win2016-gps-dm101/

The Library
UC SAN DIEGO

Data Management 101 - School of Global Policy and Strategy

Feb 17- Mar 2, 2016

12:30 pm -1:50 pm

Instructors: Juliane Schneider, Reid Otsuji, Tim Dennis, Hyeonsu Kang

Helpers: Reid Otsuji, Tim Dennis

General Information

This is the website for the School of Global Policy and Strategy short course in data management and SQL. This course will introduce you best practices in data management. In order to earn a certificate of proficiency, you must attend all of the class meetings for the course, do the short coding assignments, and pass the short, in-class quiz at the end of the course. Courses begin on Monday, January 4th. This is the first day of instruction, winter quarter.

Who: The course is aimed at GPS graduate students. **You don't need to have any previous knowledge of the tools that will be presented at the workshop.**

Where: Room 3202, 9500 Gilman Drive, #0519, La Jolla, CA. Get directions with OpenStreetMap or Google Maps.

Requirements: Participants must bring a laptop with a few specific software packages installed (listed below).

Contact: Please mail timdennis@ucsd.edu for more information.

Assessment: A certificate of proficiency will be given to students who:

- Attends all classes
- Passes all in-course quizzes
- Satisfactorily completes the weekly assignments

Need help?: Email timdennis@ucsd.edu or schedule an appointment

School of GLOBAL POLICY & STRATEGY

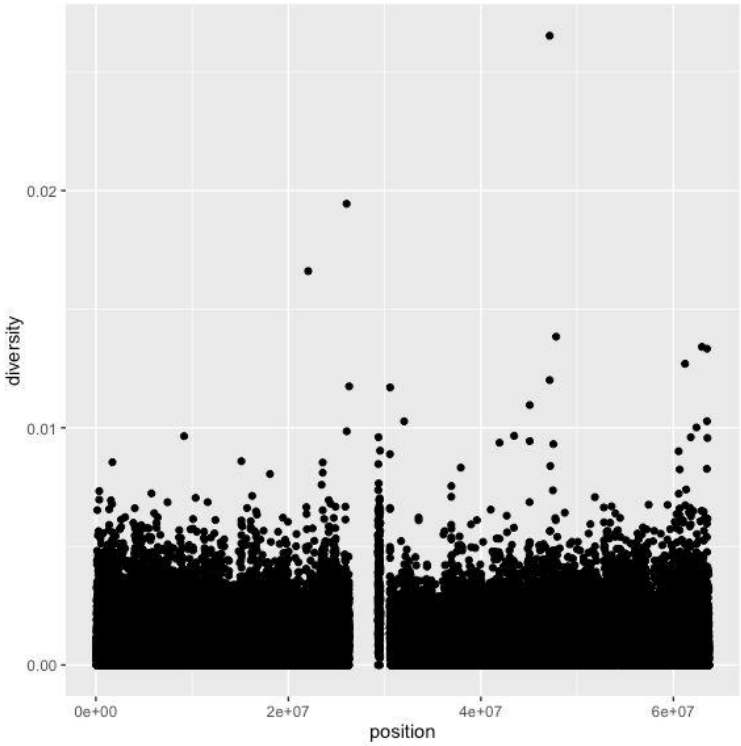
Disciplined Focused Data Instruction

R for Genomics:

- Teamed with bioinformatics librarians to teach a R 3 course series
- Courses included working with, managing and visualizing genomics data in R
- Total enrollment: 90 participants
- Future planning: offer workshop that includes High Performance Computing

http://ucsdlib.github.io/workshops/posts/intro-r-genomics/

```
d$position <- (d$end + d$start)/2
ggplot(d) + geom_point(aes(x = position, y = diversity))
```



Disciplined Focused Data Instruction

Python for Economics Department:

- 3 week Python series
- 20 Ph.D. candidates
- Courses included an introduction to Python, working with data in Python and working with web data in code

http://ucsdlib.github.io/workshops/posts/python-econ/

Challenge

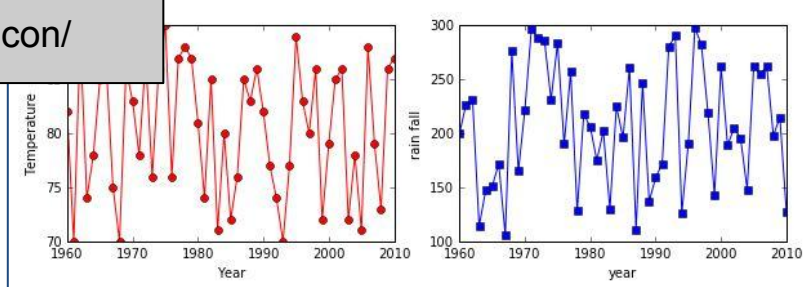
Using the data in A2_mosquito_data.csv plot the relationship between the number of mosquitos and temperature and the number of mosquitos and rainfall.

```
plt.figure(figsize=(10.0,3.0))

plt.subplot(1,2,1)
plt.plot(data2['year'], data['temperature'], 'ro-')
plt.xlabel('Year')
plt.ylabel('Temperature')

plt.subplot(1,2,2)
plt.plot(data['year'], data['rainfall'], 'bs-')
plt.xlabel('year')
plt.ylabel('rain fall')
```

```
<matplotlib.text.Text at 0x10d294080>
```



Software Carpentry at UC San Diego

Library Affiliation with SWC and Data Carpentry

Completed SWC instructor training and certification:

- Data Librarian
- Metadata specialist
- Data Curation Specialist

Software Carpentry and Data Carpentry instruction:

- Hosting SWC events - general and discipline specific
- Library Carpentry event - modified instruction for library staff
- Using open lessons for Python, R, UNIX Shell, Git, OpenRefine

Future topics of interest:

- Text Mining
- High Performance Computing
- Computational cloud computing

software carpentry

DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Library Carpentry

Software Skills Training for Librarians

The Library UC SAN DIEGO

Software Carpentry at UC San Diego

SWC instruction requires core pedagogical concepts and evidence based teaching practices.

Teaching basic computational skills for research computing.

- 2 Software Carpentry Workshops
  - Fall 2015 & Spring 2016
  - Certified Instructors form SIO, SDSC and the Library
- 2 full day workshop
  - Unix Shell
  - Programming Python
  - Version Control in Git
- 70 total participants

http://ucsdlib.github.io/2016-05-17-ucsd/

http://scicomp.sdsc.edu/2015-10-27-ucsd/

Schedule

Surveys

Please be sure to complete these surveys before and after the workshop.

Pre-workshop Survey

Post-workshop Survey

Day 1

09:00	Automating tasks with the Unix shell
10:30	Coffee
[illegible]	Lunch break
13:00	Building programs with Python
[illegible]	Coffee
16:00	Wrap-up

Day 2

09:00	Version control with Git
10:30	Coffee
12:00	Lunch break
13:00	Building programs with Python (pt. 2), Pandas
14:30	Coffee
16:00	Wrap-up

Follow-on Positive Benefits

- One-on-one consultations after workshops
- Increased engagement with researchers and their research
- Through Software Carpentry network able to develop local communities (HackerWithin)
- Opportunities for partnering with campus entities (research computing, etc.)

Data Science Student Society

THW The Hacker Within

The Hacker Within began as a student organization at the University of Wisconsin-Madison, and is now reborn as a collection of such chapters around the world.

Each of the chapters convenes a community of scientists, at all levels of their education and training, to share their knowledge and best practices in using computing to accomplish their work.

Current chapters:

- U. California, Berkeley (USA)
- U. Wisconsin-Madison (USA)
- Swinburne (Australia)
- U. of Illinois, Urbana-Champaign (USA)

Nascent chapters:

- U. California, Davis
- U. Melbourne (Australia)
- Yale (USA)
- Michigan State (USA)

Challenges

- Developing the skills in house to be able teach these new tools
- Changing the perception of traditional library
- Administrative overhead in planning and organizing workshops
- Promoting workshop events to the wider campus community
- Finding instructors
- Data science and its uses in research is dynamic

The Library
UC SAN DIEGO

Opportunities

- Library becomes facilitating factor and support service in the future of research
- "Library as Switzerland"
- SWC provides tested lessons for re-use and an active community of instructors (great way to *skill up*)
- Data instruction promotes engagement with researchers and students

The Library
UC SAN DIEGO

Thank you!

Questions?

Contact us:

Tim Dennis
Data Librarian
timdennis@ucsd.edu

Reid Otsuji
Data Curation Specialist Librarian
rotsuji@ucsd.edu

The Library
UC SAN DIEGO