Designing a Synergistic Relationship Between Undergraduate Data Science Education and Usability of Biodiversity Databases

Ciera Martinez, PhD
Berkeley Institute of Data Science
Mozilla Foundation
1. Motivation

2. Project set-up

2. Results of the project

3. End questions
1. Motivation
2. Project set-up
2. Results of the project
3. End questions
How do organisms get their shape? Evolution!
Learning to perform computational research is hard

- Attempting to gain the correct skills is frustrating
- They don’t teach the skills you need in classes
- PIs don’t teach these skills to their students

How I learned

- Using online tutorials and documentation
- Talking with peers, largely online
- Teaching others

Now committed to programming education
Motivation

Utilizing data made my research faster and better

Chitwood et al, 2014, Martinez et al, 2019, Martinez et al, 2016a, Martinez et al., 2016b,
Motivation
Motivation
I want to go to there.
Yay!

So much cool data!
Motivation

Yay! So much cool data!

Nay! It’s hard to:
- Navigate these databases
- Access data within these databases
- Understand what is in these databases

Ect
Motivation

Data Science Education + Using Biodiversity data To answer research questions =

Curiosity Data Project
1. Motivation

2. Project set-up

3. Results of the project

3. End questions
**Curiosity Data Project**

To use biodiversity databases and inquiry based learning to teach data literacy skills to undergraduate and graduate students.

**Project Overview**

Guided by mentors, students are given free rein to explore a database of their choosing with the end goal of creating a tutorial outlining their process.
Project set-up

Synergistic Relationship

- give feedback
- use data
- recruit lifelong advocates
- create tools
- create guide to use the database and handle data

[Arrows indicating exchange]

- provide approachable open data
- provide a point of contact for questions

Curiosity Data Project

Data Builders & Maintainers
1. Got team together

- Recruited eight undergraduates from pool of 30
- Some programming knowledge (R or Python)
- Majors: CS, Stats, Biology
- Recruited mentor help: 1 PhD student and 1 industry data scientist
2. The students were asked to pick one database and document how they access, clean, and analyze the data from that database.
2. The students were asked to pick one database and document how they access, clean, and analyze the data from that database.

List of prompt questions to begin exploration

• What features you find most useful?
• What kind of questions can we ask when using this database?
• What tools are available to use this database?
• What skills are required to use this database?
• How do you access the data?
• What tutorials help you access the database?
• What are the challenges to using this data?
• What are the rules for using this database?
Project set-up

3. Student end project goal is a tutorial that will be posted *online*

**Reasoning**

- We are giving back to the community by teaching others
- They get over their fears of posting their code online
- Expose them to open science philosophy and computational reproducibility
- Focus on code readability
- Allow them to learn how to present a story
- Teaches them computational research notebook management
- Regularly use Git and Github
- Start online presence and portfolio
4. We met for 20 min individually each week to discuss what they worked on, challenges, and plan next weeks work. (~3 hours a week)
I am so impressed with their work!
Ask how animals interact with network graph visualization

Yikang Li

GloBI

curiositydata.org
Tracking animal movement in time and space

Jackals in April 2009

MoveBank

curiositydata.org
Results of the project

Mapping dinosaur and plant fossil data in time and space

Paleobiology Database (PBDB)
Coming Soon: Explore 3D CT Scans using open source tools

Morphosource

curiositydata.org
Results of the project

curiositydata.org
Exploring the data of natural history

Digitizing Data at the National Museum of Denmark
I was fortunate enough to visit Copenhagen a few weeks ago and couldn't resist tracking...
2018, May 15 — 6 minute read

Movebank Part II: tracking animal movement in time and space
In my previous post, I talked about the awesome data resource that is Movebank, and...
2018, Apr 17 — 18 minute read

Movebank Part I: database summary and downloading a study
Results of the project

What worked?
What didn’t work?
What did the students like about the project?
What did the students dislike about the project?
What worked?

- They all thrived in self directed inquiry
- These tutorials can be coopted for workshops AND be remixed for lesson modules
- Feedback works if directly in contact with developer
- They all enjoyed it and stuck with it, I didn’t lose one student
- Students became extremely interested in the data
Results of the project

What worked?

- The students skills did improve!

![Boxplot showing skill improvement](chart.png)
Results of the project

What didn’t work?

- They sometimes floundered with so much freedom
- It was hard to assess how much work they did each week
- They could not create tools like I originally planned
- The limiting factor is me, hard to scale.
  - Reviewing the code
  - Dealing with notebook to website
  - Individual meetings
Results of the project

What did they like about the project?

- “Ability to take the project in whatever direction I wanted”
- “I love manipulating data programmatically and making plots with the data.”
- “I enjoyed learning new skills outside of the classroom and actually being able to create a tutorial using data that I was really interested in!”
- “Exploring different data visualization methods and tailor them to fit the data set”
What did they dislike about the project?

- “Debugging code and figuring out how to make an interesting…story out of the data”
- “Downloading data using API and storing big data.”
- “Lack of structure”
- “Trying to figure out how to solve the problems like dealing with missing and "bad" data”
- “That some visualization methods/packages are difficult to install and use “
Results of the project

What did they dislike about the project?

- “Debugging code and figuring out how to make an interesting…story out of the data”
- “Downloading data using API and storing big data.”
- “Lack of structure”
- “Trying to figure out how to solve the problems like dealing with missing and "bad" data”
- “That some visualization methods/packages are difficult to install and use “

But everything they listed is normal frustrations about working with code, so they learned what is common frustrations with the code
CURIOSITY DATA TEAM

Samantha

Zoe

Shirley

Winifred

Sara

Ciera

Lynn

Yikang

Ollie

Caryn

*Not pictured Yihui
Outline

1. Motivation

2. Project set-up

2. Results of the project

3. End questions
- Is there something here with gender and this data? Why? How do we leverage for bridging gender gap in STEM?
- Can this scale as an academic course?
- Can this scale in the online community?
- Do you need to know how to program to run a similar project?
- Will people use these tutorials?
- Where is an appropriate place to house these tutorials?
Thank You!

Neotoma, PBDB, GloBI, Movebank, Google Earth, iDigBio, BHL, Morphosource

All the museums, collections, and universities that hosted me to visit, and provided insightful discussion and critical feedback


Special Thanks: Deb Paul
Learning Objectives

- identify assumptions about data
- formulate research questions
- tell a story with data
- visualize data
- build data science portfolio
- write collaborative code
- employ version control
- apply data / file management
- think with reproducibility in mind
- handle and merge data
- clean messy data
Overlapping Data Literacy Skills

Minimum

1. Manipulation of data frames
   • subset
   • merge
   • transform long vs wide
   • apply formulas to values
2. Map points onto a geographic map, understand assumptions
3. Clean taxonomic names, understand assumptions
   1. Species concept
   2. Spelling errors
   3. Species synonyms
   4. Hierarchical classification
   5. Species databases available and differences between them
4. Time stamp data formats
   1. Time collected vs time measured
   2. Time formats
   3. Visualization techniques for time and space
5. Look into biology for sanity checks
6. Understand the value of plotting data and which visualization is important