# Graduate students' attitudes toward data sharing and reproducibility

Katie Mika
Data Services Librarian
Harvard Library & Institute for Quantitative Social Science
February 10, 2021
katherine_mika@harvard.edu

# Ithaka S+R Study

Large scale collaboration in 2017-2018 between research teams at 11 academic libraries and in partnership with American Society of Civil Engineers to “examine the changing research methods and practices of civil and environmental engineering scholars in the United States with the goal of identifying services to better support them.”

Cooper, D., Springer, R., Benner, J. G., Bloom, D., Carrillo, E., Carroll, A., Chang, B., . . .Yu, S. H. (2019, January 16). Supporting the Changing Research Practices of Civil and Environmental Engineering Scholars. https://doi.org/10.18665/sr.310885

“I do have concerns that I may not be quite living up to the expectations of the funding agencies as far as data management.”

- Ithaka S+R study

# New Research Questions

Are civil and environmental engineering graduate students learning about data management?

If so, how? From whom?

If not, where are the gaps?

Are there specific disciplinary practices that affect RDM?

# Interview Questions

What are your plans for managing your data?

What is your plan for your data after you graduate?

Do you typically produce data?

Do you create these data personally or are they “inherited?”

Do you work with data created by others?

What is a typical size of the data you work with?

What do you do to prepare your data for others to use?


How do you share your data with collaborators?

Do you publish your data?

What parts of your data do you publish?

What type of data do you work with?

What tools do you use to work with data?

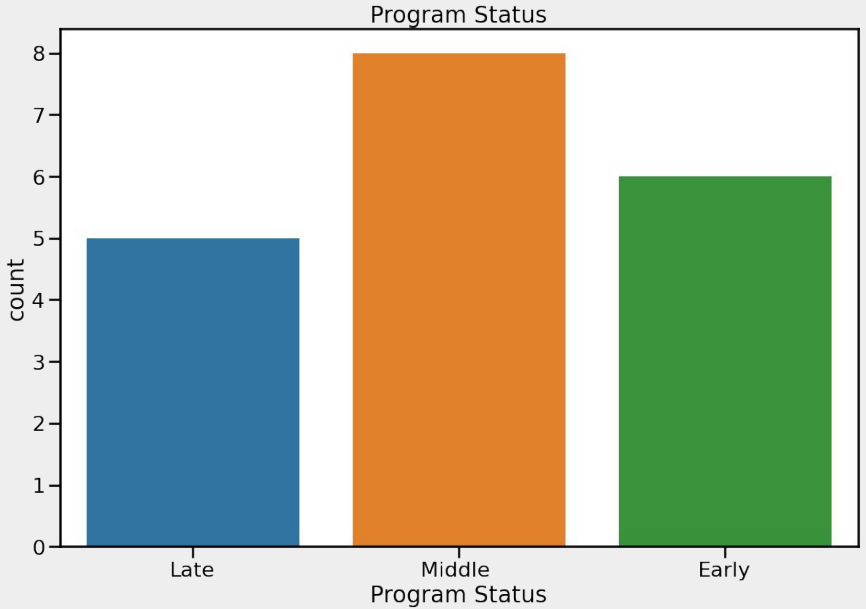
How do you backup your data?

Do you use version control?

How do you analyze your data?

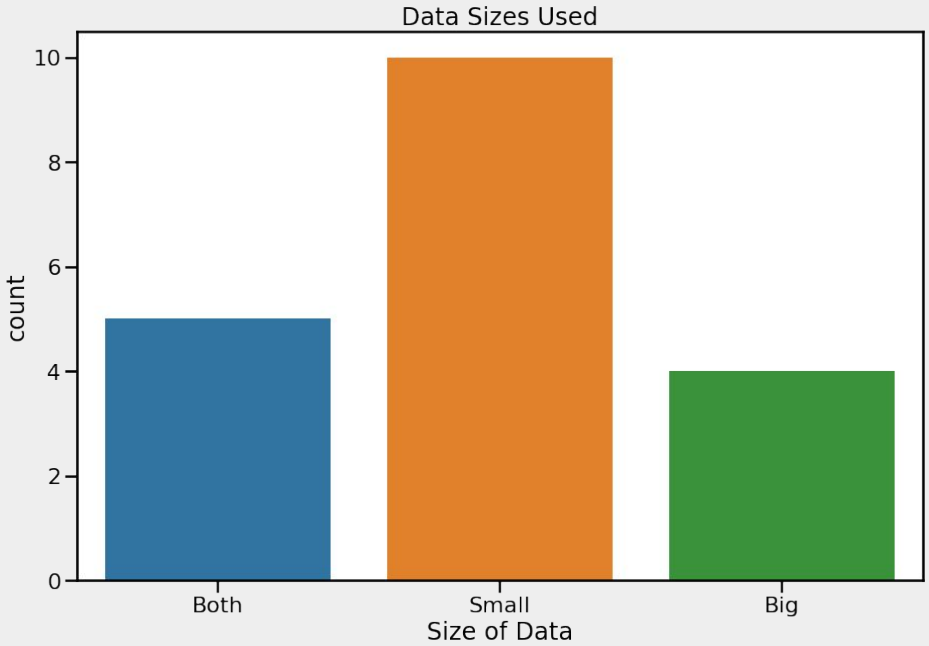
# Student snapshot

Year in Program



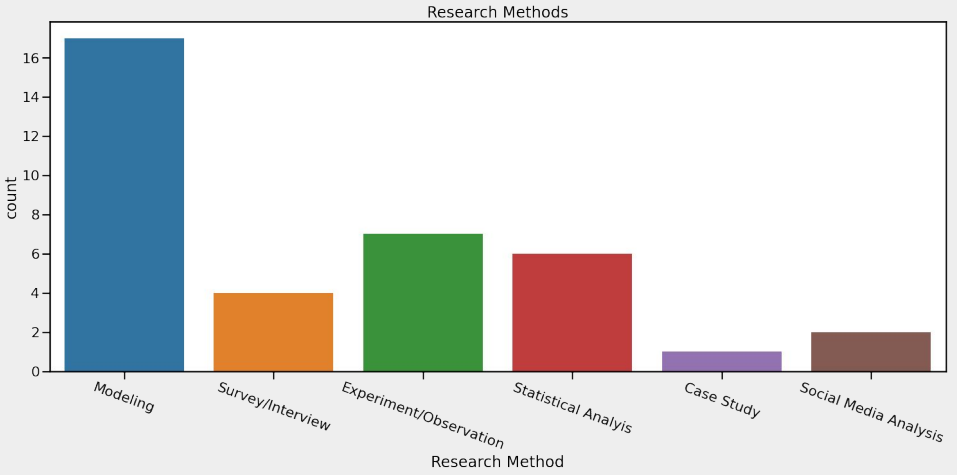
# Student snapshot

Typical size of data



# Student snapshot

## Research Methods



# Results

## Data Sharing Norms

- Sharing mostly plots, visualizations, workflows/pipelines, code & models
- Often discussed where and how they discovered data with colleagues and other students. More common to share access than directly pass around data.

# Results

## Data Sharing Norms

*“When they get to be good enough, of course, yeah. I'm still working on them. Of course if I can get to the point that I'm confident to put them online and people would benefit from them, if they can add value, by all means I can make it publicly available.”*

# Results

Data Sharing Norms

- Still seemed to value the concept of data sharing

*“An article is just text and equations, but I think nowadays what we do is really hard to capture in just text and equations. …there's a lot of hidden methods and understandings that are hidden in the code that somebody can't pick up just from the paper.”*

# Results

Reproducibility

- We did not directly ask about reproducibility, but nearly half of students brought it up anyway

*“...because even in like supplementary information, you might have all the algebra or some equations, but it's never at the detail of the code, at the level of detail of the code. So, it's useful for future students or other scientists to see how it was actually implemented. And, … often there are things that are unsaid.”*

# Themes

- Tension between wanting to publish and wanting to share openly

*“..if I'm working on a paper about energy consumption I would probably wait for that to be published and then release the data. But I think the more people sharing the better.”*

# Themes

- Sharing data is a lot of extra work

*“I was joking that like I would be such a hypocrite because I'm such a reproducibility like evangelist. And right now, I'm just working. It's probably going to take me like a week to get this thing to where it would get fully documented and reproducible. And I was like, you know what, I really could just not do this.”*

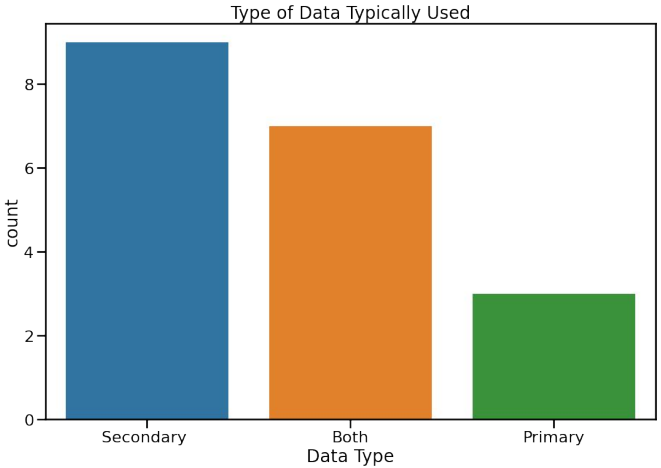
# Themes

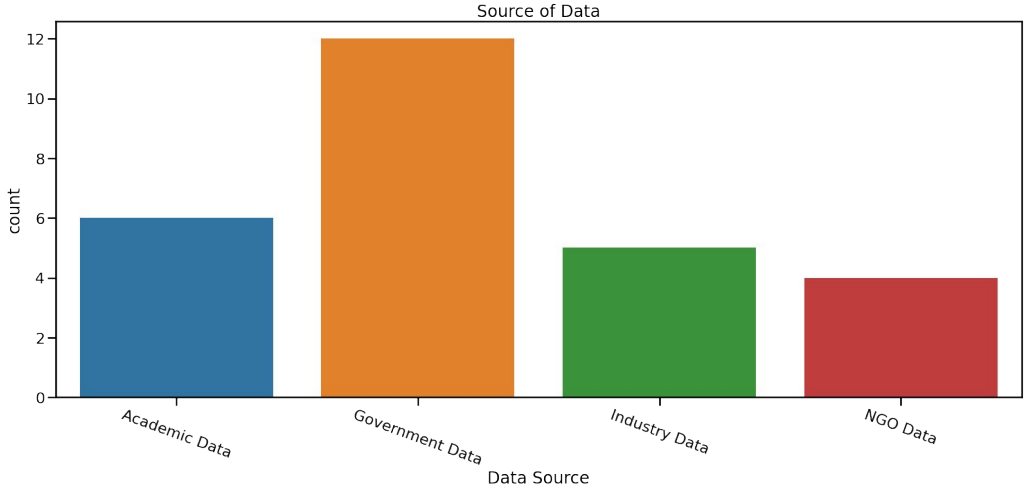
Highly generalized use of the term “reproducibility”

- Spectrum
- Can the product be reused or reapplied?
- Less closely tied to reproducing study results

# What does this mean?

- Common for CEE students to re-use data



# What does this mean?

Students seem to develop their attitudes about data sharing based on the quality of data they work with.

*“I spent a very long time trying to reproduce someone else's results, and it was just crazy. Like, **I couldn't get anything close to the results that they were** just by following what they did in the paper. I was not doubting or anything that their results were wrong; it's just that there are intermediate steps that they did not say in the paper that were actually very important to get those results.”*

# Conspicuously absent…

- No mention of funder requirements
- No mention of PI or advisor as playing an important role in learning about data sharing and reproducibility

# Conclusions, recommendations, further research

- Reproducibility-specific training is a high-value gap in graduate curricula
- Difficult secondary data may not be all bad…

- Apply this to other disciplines
- Compare qualitative results to curriculum & library instruction data
- Quantitatively measure data sharing among engineers and graduate students

# Thanks!

katherine_mika@harvard.edu

**Project Team:**

Emily Dommermuth, Science & Engineering Librarian, CU Boulder

Julie Chen, Librarian, Engineering, Carnegie Mellon University

Rebecca Kuglitsch, Lead, Branches & Services Team, CU Boulder

Abbey Lewis, STEM Learning & Collections Librarian, CU Boulder

Jessica Benner, Librarian, Computer Science & GIS, Carnegie Mellon University

Sarah Young, Senior Librarian, Social Sciences, Carnegie Mellon University

Matthew Marsteller, Principal Librarian, Engineering & Science, Carnegie Mellon University

# Results

## Basic data sharing results

- Data sharing norms
  - mostly graphs, visualizations, workflows, code & analysis, not raw data
  - Some discussion about sharing where they found data or data that they found from somewhere else
  - Wanted to clean up first or no perceived need
- Still seemed to value data sharing
  - Some specific motivations included getting information to practitioners, discovering unintended uses of your data, increasing the credibility/reputation of the research group, and their own experience as a reader not having enough information to understand the results presented in a published paper.

# Study description

Original Ithaka study

What about grad students?

Is any of this limited knowledge trickling down? (not from PIs)

Is there any coordinated attempt to teach grad students about RDM? (not really)

Are there any specific disciplinary practices or standards that impact RDM? (yes! We think!)

# Study description

# of participants

Recruitment from a variety of different labs

Description of survey

Snapshots:

- Status in program (early, mid, late)
- Type of analysis used
- Other?

# Results

Basic reproducibility results

- Didn’t ask, but still came up in about half of interviews
- Didn’t love that it was hard to reproduce or reapply results

# Discussion

- Definition of “reproducibility”
- What affects importance of reproducibility
  - Type of research
  - Subfield practices
  - Participation in academic orgs
  - Did they use secondary data
  - Data that are difficult to reuse or are irreproducible - materials are insufficient to complete a replication or reproduce results (not that the results differ) - clearly affect students’ perception of data quality and the importance of reproducibility
- Divergences from other disciplines
  - Research methods
  - Data sources
- What isn’t there (to take with a grain of salt)
  - Funder requirements
  - Any mention of advisor, PI, or class

# Recommendations

- Consider how discipline and status of researcher may affect RDM needs & practices
- For graduate students specifically:
  - Reproducibility-specific training is a high-value gap in graduate curricula that students would likely be interested in learning about
  - Have structured learning opportunities so that students who don’t have robust informal networks can succeed
  -