# STA 393 Data Science Education: From Tutorials to Assessment

**Spring 2022 Reflection**

Evan Dragich

Supervised by Dr. Mine Çetinkaya-Rundel, PhD.

The following document serves as the final artifact representing my work in STA393 this Spring 2022 semester. I met with Dr. Çetinkaya-Rundel weekly on Fridays 9-10am this semester, along with three other IS students.

## Data Science Tutorials

While the {dsbox} package was not the primary focus of my work this semester, I still oriented myself with the project and contributed two main additions.

First, I resolved a pending issue calling attention to the ambiguous levels of a particular variable. I set out to determine which numeric level corresponded to which qualitative description, and found that the levels were indeed conflated as the tutorial stood. In the process, I also updated the URL for this data's source to a more permanent link, as I was unsuccessful using the given link as-is, removing some of the HTML portion of the URL.

Then, I added the missing documentation for the Denny's and LaQuinta datasets, resolving another pending issue. In the process, I learned much about {roxygen2} and the relationship between .Rd files, .R files, and the documentation that appears in the RStudio pane when you call `?`.

Finally, the culmination of my semester's work with {dsbox} was generating a tutorial using the datasets, which is the tenth in the package so far. Modeled after the Data Science In a Box labs associated with these datasets, the tutorial serves as a capstone to the previous tutorials, integrating previously-seen topics such as descriptive statistics, data wrangling, particularly `mutate()` and `*_join()` functions, as well as spatial mapping. Through this, I gained a significant confidence using {learnr} to set exercise code chunks and provide hints, as well as {gradethis} to automatically give feedback on inputted answers. I found it very enjoyable to consider the structure of the tutorial, as I sought to balance having one overarching narrative with a variety of topics and questions to get at many facets of the data science curriculum.

While this pull request has yet to be merged, all checks are passed. Since I began work on {dsbox}, I have also updated the `DESCRIPTION` file to include all currently-used packages and render using a new version of {roxygen2}. I believe that next steps for this project consist of user-testing, to encounter both small bugs or typos and the high-level flow and user experience navigating the tutorials. All data is up to date and all checks are currently passed, so if the current iteration meets the desired specifications for such a package, I think a CRAN release is definitely in the near future (especially because the documentation is somewhat inherently accounted for in the nature of the tutorials themselves).

## Data Science Assessment

The bulk of my work this semester focused on the creation of a Data Science Assessment. When I began, the project consisted of a draft ~50 question assessment, to which I provided initial feedback These were then translated into pull requests, a new tool for me, but not before converting the entire assessment to a Quarto Book. I was completely new to Quarto, and enjoyed the practical motivation to learn another new tool, cleaning up the raw questions in the process. Some highlights of my cleaning work include: delineating items into 26 discrete passages, aligning the item numbers accordingly, ensuring that all content translated well to the HTML Book format, creating a download PDF link on the website and ensuring all content rendered well in PDF form, and adding code-styling to the pseudocode and table passage using HTML.

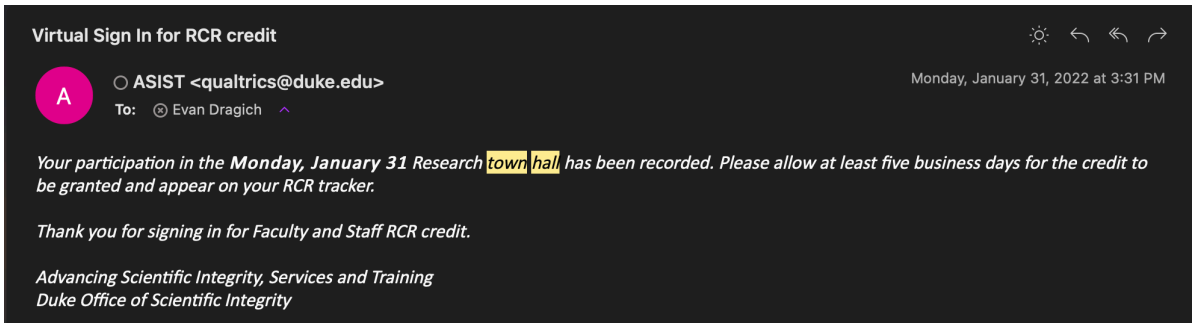
We were then ready to gather feedback on the polished assessment, and scheduled our first interview with Dr. Nick Horton, of Amherst College. The interview occurred on 4/21/2022 for 2 hours, and Dr. Çetinkaya-Rundel and I walked through the assessment with him, taking notes and eliciting interview responses, respectively. We had some immediate work after the interview, fixing a handful of "obvious" issues illuminated via the interview.

The interview audio recording and transcription file were saved to the cloud, and I downloaded into NVivo to begin coding. Following the stellar guidelines given in Reinhart et al., I believe that simply coding the "tags" that Dr. Horton gave each item as well as some general themes in the introductory and concluding open-ended questions will suffice for our rounds of informal faculty interviews. However, I have gotten good practice loading audio/text into NVivo, and playing with timestamps to draw out themes. I look forward to possibly using this tool to more closely analyze student data.

Currently, the next steps are to contact more professors to repeat the interview process, possibly pruning items beforehand. While these may not occur until the fall semester, there is a chance that work will continue over the summer to be able to gather student data as soon as possible.

## Responsible Conduct in Research Requirement

The final portion of my required work as a student in a research independent study is attendance at, and reflection upon, one Duke-sponsored Responsible Conduct in Research RCR event during the enrolled semester. I attended the 90 minute "Academic Research and Social Responsibility" research town hall held on January 31, 2022.




The RCR requirement also entails completing the three required RCR Basics courses and one additional elective. I had previously completed those for another research experience; see the attached pdf at the end of this document.

The following summarizes each of the town hall's speakers' topics and connections to my work:

First, Dr. Biederman led with a quote describing a guiding principle of ethics in research: "nothing without us, about us." She noted that this inclusion should span the entire timeline of the research project, ideally engaging with population of interest community members before research begins all the way until after the conclusion. Her work helped build the evidence base for medical respite care. She recalled helping unhoused persons, the focus of her research, successfully log onto Zoom and conduct the interview. This was a reminder of how mucch I take the convenience of Zoom for granted when we are planning the faculty interviews for this project; the faculty demographic, particularly STEM faculty, that we are interviewing make for very easy gathering of data. The other portion of her talk that stuck with me was her note that, for this study, *everyone* involved was paid the same stipend–researchers and unhoused subjects alike. Just as she noted, all parties come in with a different expertise, all of which is valuable.

Then, Dr. Rose began explaining the role of higher education in empowering marginalized and excluded groups in the US. Particularly, access to higher education turns out to play a large role in shaping political landscape and developing engaged citizens, the latter of which is the main focus of her research. First, she noted that the gender gap in bachelor's degrees awarded in the US started small iniitally, expanded to favor men around WWII, until the 1908s when women took over and that trend has continued to this day. Why? Changing demographics, changing labor force that have brought women into workforce, and womens' rights movements, mostly. She clarified there isn't much of role of public policy; that's her work that tries to

tease out whether any policy changes have been made or if it is a coincidence. Similarly, she noted that a high percentage of HBCU alumni are now serving in landmark elected official roles. Is there something particular to the HBCU experience driving this elite level of political engagement, she asks?

To explain these phemonena, she has developed policy feedback theory: the idea that by providing the income support via scholarship to facilitate pursuing a degree, the necessary organic encounters with the political system and elected officials may spur someone to engage more politically. She offered an example of her own policy lab in Sanford, where they reach out to lawmakers to see if introductory policy research by undergrads could help any NC lawmakers to further facilitate this connection.

While I wasn't sure how much the topic of informed citizenship would pertain to my data science assessment research, the idea of studying what motivates and captivates people to engage is certainly relevant to designing introductory data science curricula. Rather than just drill in statistical concepts and rote functions, the goal of any data science curriculum should be to inspire students to be lifelong learners and continue using the tools they've learned to expand their knowledge, regardless of their passions or fields.

Next up was Dr. Glymph, who began with a map of slave camps from the US Civil War period. She posed the question: Why does writing about a humanitarian crisis where victims are Black seem like advocacy? Lots of people who attend presentations of her research say "sorry," more specifically "I'm sorry about your people," but she noted that those folks get surprised when she simply says "its American history".

This led to her asking herself: Why does this archive seem different to different people? Why is this phenomenon different than American Revolution archives, say? Thus motivates her work, navigating the tension of social responsibility as a scholar, or of rearchiving US history.

"History is more than a mass of details," she posits. But it's also not advocacy in the way people think of. Historians recognize bias in the archive, like access to past documents not being unproblematic, nor are the documents themselves, whether they be testimonials from enslaved people or enslavers. She then brings up peer review, noting that we subject our work to professional standards, and we understand we cannot know the past in all its completeness and complexity, but we have to work harder to incorporate other people. She concluded with a slide of a photograph taken during the civil war of a refugee camp that housed women and children primarily, a stark contrast to the brutality displayed in the first slide. This represents the kind of work that she uses to rewrite the story of the civil war, studying several hundred camps of this kind.

While I am fortunate my research does not probe such ethical boundaries, it is still important to remember the social responsibility as a researcher when conducting work. For example, I would want to make sure that any work on student attitudes or confidences with the data science assessment properly accounts for stereotype threat, as it may not just suffice to gather feedback from historically marginalized groups in addition to dominant ones, but rather to actively recruit and focus the data collection on these underrepresented groups.

The next speaker, Dr. Mervin-Blake, echoed much of the above, citing a statistic that in clinical and translational research, 83% of participants identify as white compared to 5% Black and 1% Latino. In general, she says, there is limited access to trials and information about them, let alone meaningful community partnerships and engagement. The research is conducted, but the results must be disseminated to the patients and communities. As hopefully many know, there is a pattern of historical and current atrocities in research and the healthcare system sustains this lack of trust. Consequences arelack of diversity in the workforce, participant recruitment plans rarely include diversity recruitment strategies, POC not invited to participate in clinical trials at the same rate as their white counterparts. Again, we hope to actively include members from historically marginalized groups in our research, as the ethics component is huge in data science. In fact, a handful of current items in the assessment reference systemic biases in AI and ML and the need for deidentified data, on top of striving to be as accessible and least jargon-y as possible.

I unfortunately missed the majority of Dr. Murray's talk, but she focused on how human societies understand, value and interact with the non-human world around them.

Finally, Dr. Gibson-Davis concluded with a case study of identifying and preventing fraud in survey research: a study that went spectacularly wrong (in her words.) She and a team designed a multisite survey of families' experience with COVID, and even took multiple steps to prevent fraud, but in the end, 75-85% of responses were invalid.

She denoted the 11 steps used to screen out responses, some of which are reCAPTCHAs, inconsistency in answers, survey time/length, IP address, and phone numbers, and concludes that manual fraud was determined to be the culprit. This is when a participant either doesn't honestly answer questions, takes it validly once and then again a few more times, or pretends they meet criteria to take it. She laments that these phenomena have been amplified by survey advertisement websites.

She noted the primary fault was having a non-individualized link, and by offering compensation directly following completion, every single competion (instead of entry into a raffle, etc.) She said also that reporting results more regularly may also attenuate fraud, by creating more a buy-in for the survey's purpose.

She concludes with a few more thoughts to ponder. First, there are even ethical issues regarding classifying a fraudulent response. It is not so black/white as we may think! Then, she mentioned the relationship with IRB: what if responses pose potential harm to researchers? Finally, she lamented how to explain to funders that 75-85% of their funds went to paying people who don't exist.

While Dr. Gibson-Davis's results were certainly unfortunate, they offer some key takeaways for our project. First, it is necessary to unlikely we will obtain many fraudulent responses via faculty interviews, as they are all likely quite inherently motivated to provide assessment feedback. However, it will be important to think how to recruit students and ensuring best practices to elicit honest responses. Much of this is explored in the above mentioned Reinhart

et al., who conducted many such interviews, but keeping a healthy caution for potential error sources will be very valuable as we collect more data.

# COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

## COMPLETION REPORT - PART 1 OF 2
## COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Evan Dragich (ID: 9396378)
- **Institution Affiliation:** Duke RCR (ID: 1806)
- **Institution Email:** evan.dragich@duke.edu

- **Curriculum Group:** Undergraduate Student Responsible Conduct of Research Course
- **Course Learner Group:** Undergraduate Student Responsible Conduct of Research
- **Stage:** Stage 1 - Basic Course
- **Description:**

- **Record ID:** 41049208
- **Completion Date:** 18-Feb-2021
- **Expiration Date:** 18-Feb-2024
- **Minimum Passing:** 80
- **Reported Score*:** 93

| REQUIRED AND ELECTIVE MODULES ONLY | DATE COMPLETED | SCORE |
|---|---|---|
| Research Misconduct (RCR-Basic) (ID: 16604) | 18-Feb-2021 | 5/5 (100%) |
| Data Management (RCR-Basic) (ID: 16600) | 18-Feb-2021 | 5/5 (100%) |
| Duke Translational Omics Research Misconduct Case (ID: 20198) | 18-Feb-2021 | No Quiz |
| Research, Ethics, and Society (ID: 15198) | 18-Feb-2021 | 4/5 (80%) |

**For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.**

**Verify at**: www.citiprogram.org/verify/?kc0e65d21-1321-4d0a-b8a1-b73b035f0858-41049208

**Collaborative Institutional Training Initiative (CITI Program)**
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

# COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

## COMPLETION REPORT - PART 2 OF 2
## COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Evan Dragich (ID: 9396378)
- **Institution Affiliation:** Duke RCR (ID: 1806)
- **Institution Email:** evan.dragich@duke.edu

- **Curriculum Group:** Undergraduate Student Responsible Conduct of Research Course
- **Course Learner Group:** Undergraduate Student Responsible Conduct of Research
- **Stage:** Stage 1 - Basic Course
- **Description:**

- **Record ID:** 41049208
- **Report Date:** 11-Jan-2022
- **Current Score**:** 93

| REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES | MOST RECENT | SCORE |
|---|---|---|
| Research, Ethics, and Society (ID: 15198) | 18-Feb-2021 | 4/5 (80%) |
| Duke Translational Omics Research Misconduct Case (ID: 20198) | 18-Feb-2021 | No Quiz |
| Data Management (RCR-Basic) (ID: 16600) | 18-Feb-2021 | 5/5 (100%) |
| Research Misconduct (RCR-Basic) (ID: 16604) | 18-Feb-2021 | 5/5 (100%) |

**For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.**

**Verify at**: www.citiprogram.org/verify/?kc0e65d21-1321-4d0a-b8a1-b73b035f0858-41049208

**Collaborative Institutional Training Initiative (CITI Program)**
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org