A Pedagogical Approach to Create and Assess Domain Specific Data Science Learning Materials in the Biomedical Sciences

Daniel Y. Chen, MPH

Research Proposal submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

Dr. Anne M. Brown (Chair)
Dr. Alexandra L. Hanlon
Dr. David M. Higdon
Dr. Stephanie N. Lewis

February 19, 2021

Title: A Pedagogical Approach to Create and Assess Domain Specific Data Science Learning Materials in the Biomedical Sciences

Project Summary

In 2020, 89% of all hospitals have implemented an electronic health record system creating 2,314 exabytes of new medical data since the Health Information Technology for Economic and Clinical Health (HITECH) Act's Electronic Health Records - Meaningful Use (EHR-MU) clause of 2009 (Stewart 2020; Moriarty 2020; U.S. Dept. of Health and Human Services 2017; Office of the National Coordinator for Health Information Technology (ONC) 2020; Office for Civil Rights 2009). This sheer volume of health data necessitates the understanding, accessing, managing, and interpreting of data across researchers, clinicians, and patients (Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care 2010). While EHR systems have their own data challenges, the influx of electronic data has called for changes in how clinicians undergo training to meet the challenges of evidenced-based medicine by using these data (American Medical Association 2021; Bresnick 2015). By contextualizing and democratizing data science skills for clinicians, we can provide them more capacity to explore and make better use of the data (Kross et al. 2020). Additionally, by empowering those in or interested in a biomedical profession with better data literacy and data science skills, we can expand the workforce needed to better use and collect the data we need to innovate and progress health care. We can accomplish this by teaching learners the programming tools used for data analytics (Farrell and Carey 2018). This proposal seeks to address the following knowledge gaps in the literature and needs in the field of training biomedical professionals: (1) There are no formal learner personas for the biomedical community and the assessment tools to identify and create learner personas do not exist. (2) Data science learning materials for the biomedical sciences lack community oriented, open, and maintained lessons targeting learner persona needs grounded in pedagogical practices and theory. (3) While we know a lot about the teaching and pedagogy of computer science education, less is known about data literacy education, and almost nothing is known about data science education in an applied domain (e.g., biomedical sciences).

We hypothesize that learning materials with an eye towards tidy data principles is an effective way to teach the data science and data literacy skills that will help learners incorporate programming and data science skills from their spreadsheet workflows. We will be using a series of longitudinal surveys along with a data science curriculum that we will create to test this hypothesis. This work will bridge the skills gap between medical practitioners and domain experts in the biomedical sciences with the analysts, researchers, and data scientists to make better use of data (storage, FAIR, stewardship) in data science teams by creating a computational community of practice that can enhance workforce development, modernize the data ecosystem, work with data science tools for sustainable and open science.

<u>Specific Aim 1</u>: **Identify learner personas in the biomedical sciences by creating and validating learner self-assessment surveys.**

1.1: Learner self-assessment survey asking questions about prior programming, statistics, and data knowledge will be used to create learner personas.
1.2: Validate learner self-assessment survey.
1.3: Personas will encompass a student's prior knowledge using survey data. General background, perception of needs, and special considerations will be added to make each learner persona a complete character.

<u>Specific Aim 2</u>: **Create an effective data science for biomedical science curriculum based on best education and pedagogy practices.**

2.1: Learning objectives focused on core data literacy principles in the data science pipeline will be used for each lesson module.
2.2 Lesson content follow best educational and pedagogical practices.
2.3 Assess the effectiveness of learning materials.

<u>Specific Aim 3</u>: **Assess the effectiveness of formative assessments in learning objectives.**

3.1: Implement an experiment for conducting formative and summative assessment question types.
3.2: Assess the effectiveness of targeted feedback in auto-grading systems used in formative and summative feedback.

A. SIGNIFICANCE

Importance of the Problem to Be Addressed. 2,314 exabytes of new medical data was projected to have been produced in 2020 (Stewart 2020). This sheer volume of health data necessitates the understanding, accessing, managing, and interpreting of data across researchers, clinicians, and patients (Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care 2010). By democratizing data science skills for clinicians and other biomedical professionals, they will be able to better understand their patient population, better communicate with research teams to improve the outcomes of patients, and be better advocates for their patients. However, existing data science learning materials in the medical and biomedical sciences lack one of the following features: (1) is community oriented, (2) has an open creative commons license, (3) is maintained, (4) is accessible, (5) follows education and pedagogy best practices to target learning objectives, and (6) is domain specific. These are features that would modernize the biomedical data-resource ecosystem, promote Findable, Accessible, Interoperable, and Reusable (FAIR) principles, and enhance the data science and research workforce in the biomedical sciences.

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 included the concept of Electronic Health Records - Meaningful Use (EHR-MU), which incentivized all medical records to be electronic by 2014 (Office for Civil Rights 2009; U.S. Dept. of Health and Human Services 2017; Office of the National Coordinator for Health Information Technology (ONC) 2020). Currently, in 2020 more than 89% of all hospitals have implemented an EHR system (Moriarty 2020). While EHR systems have their own data challenges, the influx of electronic data has called for changes in how clinicians undergo training to meet the challenges of evidenced-based medicine by using these data (American Medical Association 2021; Bresnick 2015). By contextualizing and democratizing data science skills for clinicians, we can provide them more capacity to explore and make better use of the data (Kross et al. 2020). Additionally, by empowering those in or interested in a biomedical profession with better data literacy and data science skills, we can expand the workforce needed to better use and collect the data we need to innovate and progress health care. We can accomplish this skills expansion by teaching learners the programming tools used for data analytics (Farrell and Carey 2018).

Programming courses are generally inaccessible for someone with a different domain base with high dropout rates and a steep learning curve (Ogier et al. 2018; Farrell and Carey 2018). Motivation and mindset are some of the integral roles in learning programming and building life-long learners (Ambrose et al. 2010). A backward design approach using learner personas for creating lessons help keep teaching focused on objectives and help cater the needs of the learner to the overall learning objectives (Wilson 2019). The prevalence of Excel as a data tool guided us to focus on how spreadsheets fit in the data science pipeline and how data literacy concepts, particularly the concept of "tidy data," can be taught using spreadsheets. Data science tools are built around "tidy data" principles, a core data literacy topic describing how the rows and columns of a data set need to be specified for analysis. Once the lessons are created, it can be freely shared (e.g., using a CC-0 creative commons license) and improved upon, and has the flexibility to be adapted to individual instructor needs.

There is emphasis and need to center materials for individuals in biomedical fields that center around tidy data principles because the data science process requires a tidy dataset to begin the cycle of understanding the data before results can be communicated (Figure 1).

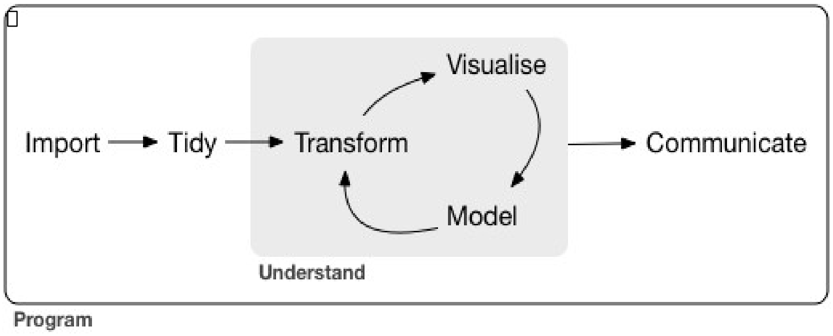

Figure 1. Standard workflow for data science project. A data science project has a feedback loop between transforming, modeling, and visualizing data before insights can be communicated. Figure taken from the R for Data Science (Wickham and Grolemund, 2016).

Unfortunately, the process of understanding and drawing conclusions from data is not this linear and requires many smaller feedback loops to account for biases and to tell a more accurate story for a decision (Figure 2). Notably, data science products usually end up with some decision or action that will affect the world. This makes each step of the data science process influential to the final set of decisions. Notably, each step of the data science pipeline is a data set, and the data literacy skills needed to process and work with the data in each step is paramount to the final results.

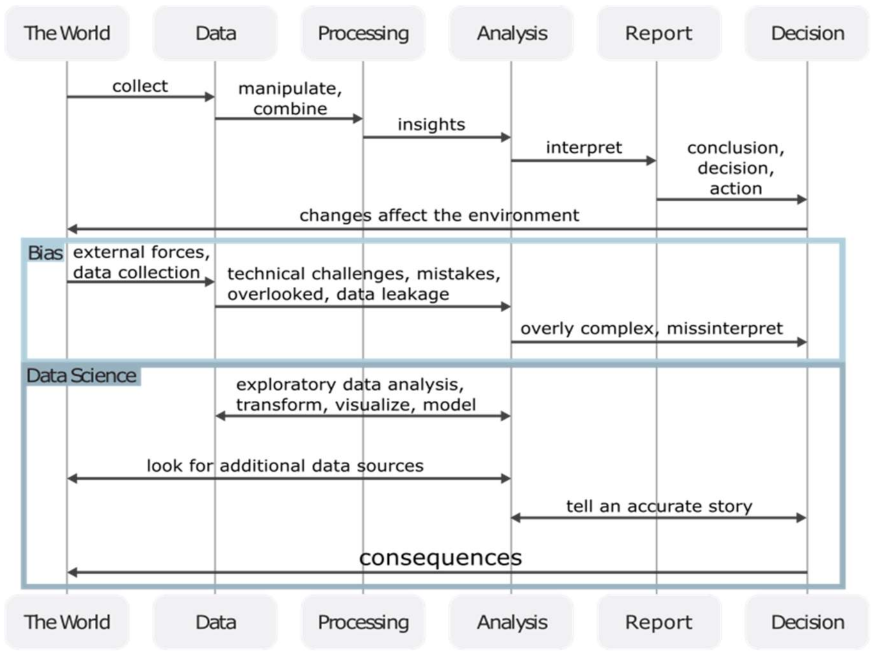

Figure 2. Overview of the data life cycle in the research ecosystem. There are many smaller feedback loops between each step in the data science process that affect the final decision, which relate to consequence in the world. This is especially relevant in the biomedical/medical domain.

This proposal seeks to address the following knowledge gaps in the literature and needs in the field of training biomedical professionals: (1) There are no formal learner personas for the biomedical community and the assessment tools to identify and create learner personas do not exist. (2) Data science learning materials for the biomedical sciences lack community oriented, open, and maintained lessons targeting learner persona needs grounded in pedagogical practices and theory. (3) While we know a lot about the teaching and pedagogy of computer science education, less is known about data literacy education, and almost nothing is known about data science education in an applied domain (e.g., biomedical sciences).

Rigor of Prior Research Supporting the Aims.

Aim 1: Identify learner personas in the biomedical sciences by creating and validating learner self-assessment surveys. Personas are detailed fictional characters based on well-understood and highly specified data to facilitate user-centered design (Pruitt and Adlin 2006; Zagallo et al. 2019). Learner personas encompass a learner's general background, prior relevant knowledge, perception of needs, and special considerations (Wilson 2019). These personas can be used along with a backwards lesson design method to keep teaching focused on learning objectives, and keep assessment materials within the scope of the learning materials (Wilson 2019). To identify the learner personas, adaption of questions from The Carpentries (Jordan et al 2018; Jordan 2016; Jordan et al 2017a; Jordan et al 2017b; Jordan 2018; Jordan et al 2020), "How Learning Works" (Ambrose et al. 2010), and "Teaching Tech Together" (Wilson 2019) and focused on 3 knowledge domains: programming knowledge, data knowledge, and statistics knowledge will be created. This learner self-assessment study will be critical in determining who will engage in this material, what needs exist in the current spectrum of knowledge, and avenues to deliver content and competencies. Personas will be crafted based on the 3 knowledge domains in data science and will be sent out to list serves and results can be clustered to identify personas using hierarchical clustering (Zagallo et al. 2019). **The personas created can help future educators in the biomedical sciences teaching data science skills focus their content, so they are relevant to the population and address their needs. The survey and persona clustering methodology can be adapted and utilized to create data science materials for other professional domains.**

Previous studies and preliminary data highlight the ability of clustering to identify personas (Figure 3). The identified clusters were combined with the original survey data to fill in each persona's prior relevant knowledge and background. The perception of needs and special considerations were created to make each persona complete but not based on survey data. A future qualitative study would be needed to get a more accurate background, need, and special considerations for the personas (Zagallo et al. 2019). Preliminary data also suggests that the survey is internally consistent and valid. However, a larger sample size across a wider geographic area would be needed to externally validate the survey.

Aim 2: Create an effective data science for biomedical science curriculum based on best education and pedagogy practices. Create a data science curriculum for the biomedical sciences using a backwards design approach. This puts the learning objectives, formative and summative assessment questions at the forefront of the lesson material to keep them focused and in the scope of the lesson (Wilson 2019). Learners who want to learn how to perform data analysis, typically, also need to learn data literacy skills to learn how to obtain and manipulate data (Milo 2005). Tidy data principles will be employed as the guiding concept of data literacy (Wickham 2014) to focus our learning objectives. Best practices on education and pedagogy dictate small, focused lessons and reinforce the learning objectives by creating a series of formative assessments (Wilson 2019; Ambrose et al. 2010). To test the content and effectiveness of the materials and its learning objectives, we use a series of pre-workshop and post-workshop surveys to determine learner's confidence in the learning objectives (Jordan et al 2018; Jordan 2016; Jordan et al 2017a).

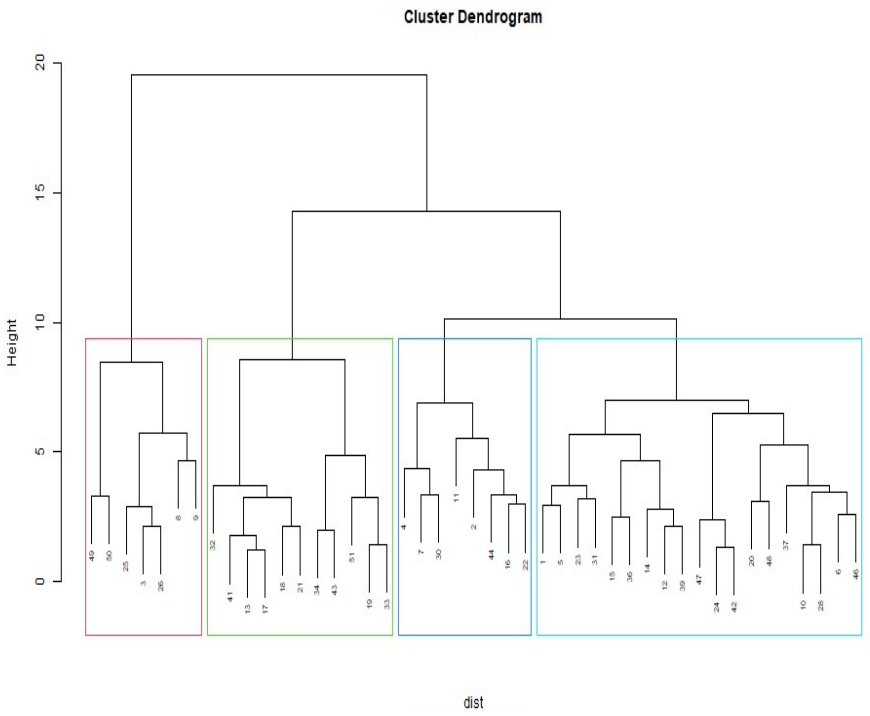

Figure 3. Learner self-assessment clusters. Four clusters created from the hierarchical clustering using Euclidean distance and Ward's method as based on preliminary data of the learner self-assessment survey. From left to right: experts (red), clinicians (green), students (blue), and academics (teal).

A long-term survey will be sent out to respondents to see how their confidence with the same set of learning objectives change over 6 months, to see how learners may have retained and built on the knowledge from the workshop. There is a final summative assessment question in both the post-workshop and long-term survey. **This work will serve as the first set of (1) community-oriented, (2) open with a creative commons license, (3) accessible, (4) follows best pedagogical practices, and (5) domain specific surveys and learning materials.** Maintainability needs to be accessed over longer periods of time, but organizations like the carpentries provide a community and mechanism where these materials can be migrated to after the initial curriculum assessment is complete to find other lesson maintainers in their incubator and lab community lessons. The surveys will be published to be used in other workshops and adapted to other domains.

Aim 3: Assess the effectiveness of formative assessments in learning objectives. Formative assessments are a pedagogical tool that instructors use to identify learner's misconceptions (Wilson 2019). In order to reduce the cognitive load on the learners, various types of assessment questions can be used. Parson's problems take a block of solution code, scramble the order of the lines, and ask the learner to assemble the code back into the correct execution order (Wilson 2019). Faded's examples provide working code snippets with some amount of the code "blanked out" (Wilson 2019). Parson's problems allow the learner to focus on the overall steps and flow of the thought process, and Faded's examples focus the learner's attention to a specific part of the code. Both provide some kind of scaffolding mechanism for the student, so they are not writing code from scratch. **These assessments will look for time to complete and solution correctness as a measure of meeting the learning objectives. It will also be the first set of data science specific formative assessments focused on data literacy topics, and not basic programming concepts in the computer science literature.** By using the learning materials from Aim 2 and the assessment tools focused on learning objectives we will create a summative assessment question. We expect to have better learning outcomes in the learners when concepts are reinforced with formative assessment questions that guide the learner to aspects of the code that are incorrect, rather than simply telling them the solution they provided is incorrect (Figure 4).

<u>Significance of the Expected Research Contribution:</u> Upon successful completion of the proposed studies, we expect our contribution to be a framework of how to create domain specific data science learning materials. The learner personas developed in Aim 1 will be used when teaching data science to new learners in the medical and biomedical sciences. The surveys used to create the personas can be used to other domains which can inform instructors about their learners. In addition, we are performing one of the few studies to date that look into how students learn in a data literacy and data science context, in connectivity to an applied field, not in a computer science context. **This contribution is expected to be significant** because of the growing need in the biomedical workforce for data science education. We are not only creating data science learning materials following best education and pedagogical practices, but also creating a curriculum in the biomedical sciences domain along with the tools and framework for expanding the content to other domains. Additionally, formative assessment questions will be measured for their effectiveness in learning the data science and data literacy contexts.

B. INNOVATION

Our work creates and validates a survey that can be used in the biomedical science to create learner personas. We have adapted survey questions from other educators to create 4 surveys: self-assessment, pre-workshop, post-workshop, and long-term workshop. These surveys are general enough to capture data literacy, programming, and statistics knowledge, while also being domain specific and flexible to be adapted to other domains. Surveys will be validated so they can be used for further studies and as a tool for educators and lays the groundwork for more survey external validation to identify data science learner personas. The surveys are used to create learner personas which are the first set of published personas for learners in the biomedical sciences, and the methods used can be used to create learner personas for other domains and other subgroups of data science (e.g., statistics literacy, data management literacy).

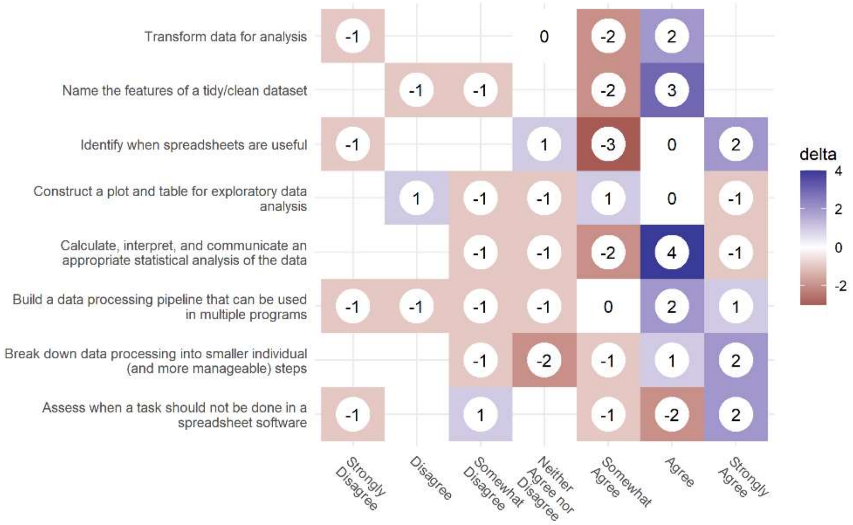

Figure 4. Influence of workshop training on pre- and post - assessment abilities. The numbers show the differences from the pre-workshop counts, from the post-workshop counts. White (0) means there was a net 0 difference between responses. Blue represents where more responses went to after the workshop and red represents the number of people who migrated from a particular response after the workshop.

Theses learning materials link the data literacy skills to the overall data science process. Many resources around data science mainly focus on the actual model fitting and evaluation of the data science process (Kross et al. 2020). Others that focus on data processing focus on discrete steps without incorporating the overall data literacy concepts. The content we have created using a backwards design approach always frames key data science steps in the context of data literacy and data processing pipelines. This creates a more holistic set of topics that are taught at the point of need, while highlighting avenues for further learning. In addition, the materials created are one of the few that are community oriented, has a creative commons license, accessible, and follows pedagogical best practices that clearly displays target audience and learning objectives.

Our experiments will teach us more about learning data science and data literacy skills, not simply programming and computer science concepts. A majority of literature centered around computer science education is focused on the programming and topics used in computer science classes. The formative question types used in computer science education informs the types of questions used in data literacy and data science, however, little is known about what formative question topics inform learning objectives in data science curriculum.

C. APPROACH

C1. Human Subject Research. We have an approved IRB (#20-537) that outlines the minimal risk from the survey participants, and have a data plan for storage, anonymization, and sharing.

C2. Introduction. While there is a lot of literature and studies on computer science education, very little is known about data literacy education, and even less on data science education. Since data science skills inherently involve programming (Song and Zhu 2016; Dolgopolovas and Dagienė 2021; Farrell and Carey 2018), there are synergies between the educational and pedagogical approach to teaching data science, however, little is known about what key concepts are needed to engage into the larger data science process and what learning objectives that need to be taught, and the effectiveness of those learning objectives.

This work will create community-oriented, open, maintained, and focused data science learning materials for the medical and biomedical sciences. To this end, in Aim 1, we lay the groundwork for understanding potential learners by identifying and creating learner personas, fictional characters that represent a typical type of learner, by creating and validating a set of self-assessment surveys. In Aims 2 and 3, we create and assess the effectiveness of the learning materials, the workshop that teaches the materials, and the implementation of formative and summative assessment questions to see if learning objectives are met. As a collective whole, we will better understand the needs of our learners and create a data science curriculum that meets their needs while providing a solid data literacy foundation that can be used to continue learning (Farrell and Carey 2018).

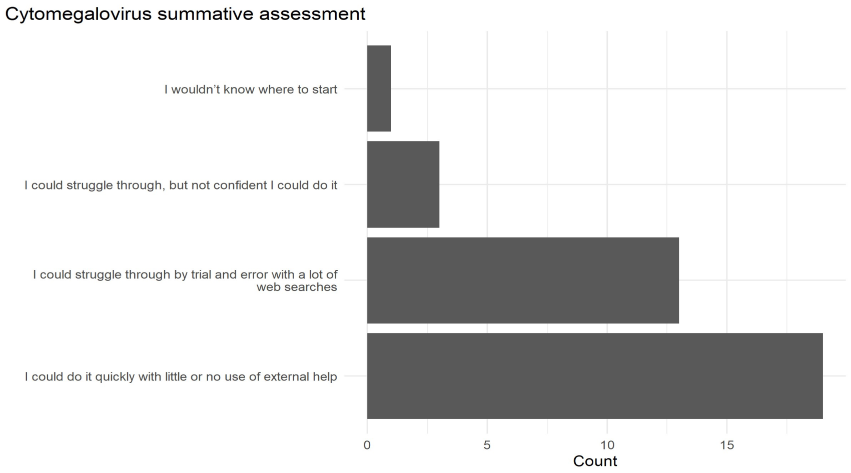

Figure 5. Example summative assessment response in post-assessment survey. The question asked the learners about a learner's comfort and ability in loading a tabulated dataset, cleaning the data, and performing a statistical analysis to answer a question on the topic of cytomegalovirus.

The long-term implications of this work are to bridge the skills gap between medical practitioners and domain experts in the biomedical sciences with the analysts, researchers, and data scientists to make better use of data (storage, FAIR, stewardship) in data science teams by creating and bolstering a computational community of practice that can enhance workforce development, modernize the data ecosystem, work with data science tools for sustainable and open science.

C3. Hypothesis. Our central hypothesis is that learning materials with an eye towards the learner and tidy data principles is an effective way to teach the data science and data literacy skills that will help learners incorporate programming and data science skills from their spreadsheet workflows. Data science tools are built around inputs that are defined by tidy data principles. Spreadsheet programs make it easy to treat data sets as a visualization, which makes the data less flexible for multiple uses. It is possible programming may not be incorporated by learners, but these materials may help curate better datasets that can be used in data science teams. To address these critical gaps in knowledge, we will create a set of surveys that will inform us of the potential learners and assess the effectiveness of the learning materials. Lesson efficacy will be tested against learning objectives.

C4. Experimental Design.

Aim 1: Identify learner personas in the biomedical sciences by creating and validating learner self-assessment surveys. Assessing the prior knowledge of potential learners in the medical and biomedical sciences who are interested in learning data science skills by creating a learner self-assessment where participants rate their own comfort in data, statistics, and programming skills.

Working hypothesis: We hypothesize that learners will fall across the 3 main groups of the Dreyfus model of skill acquisition: novice, intermediate, and expert. These groups will be distinguishable based on their own comfort in 3 domains of data science (data, programming, and statistics knowledge). To test this, we will create a learner self-assessment survey. The survey will cover data, programming, and statistics knowledge and will have at least 2 questions asking about the same underlying concept for internal consistency. Results from the survey along with demographic information will be combined to create the personas. These personas will be used to inform the learning objectives for lesson materials.

Preliminary Data for Aim 1. In preliminary studies, we found that we were able to cluster the respondents into 4 clusters using hierarchical clustering with Euclidean distance and Ward's method. We then combined these clusters with the occupation question to come up with the 4 learner personas: clinicians (novice), academics (intermediate), students (intermediate), and programmers (experts). This gives us the framework we can use to expand the study to a larger population to both validate the survey instrument and the results.

1.1 Learner self-assessment survey asking questions about prior programming, statistics, and data knowledge will be used to create learner personas. We will send out self-assessment surveys to various medical and biomedical groups around the Virginia Tech campus. We collected preliminary data over the summer of 2020 and had 51 participants consented to the survey. Figure 6 shows the grouped distribution of responses to the occupation demographic question (this is a select-all-that-apply question).

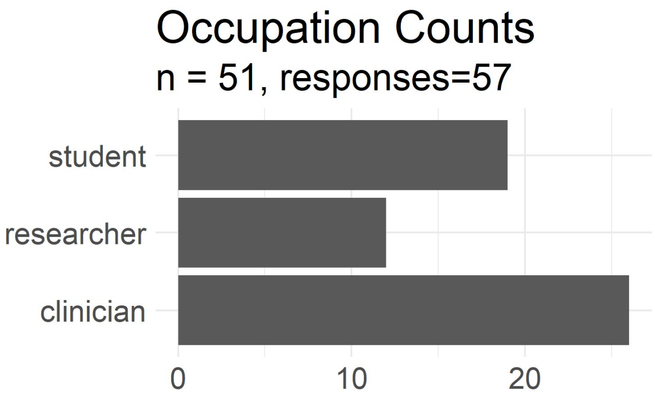

Figure 6. Grouped demographic counts. Counts of the occupations from learner self-assessment survey aggregated into one of the 3 groups shown. Each respondent had the option to select more than one occupation that applies to themself.

In general, we found that the overall group has low programming skills with basic data analysis skills primarily using Excel. They do not understand how data pipelines are created, and do not know how data can be processed into different "shapes" for analysis. An important framing of the materials would be to start with spreadsheet programs and tie their use into more advanced skills using programming languages (Farrell and Carey 2018). The survey also asks a summary Likert scale table of questions (Figure 7). These results confirm the overall findings where respondents typically do not use a programming language in their work and are indifferent towards programming in doing analysis. They did report that having access to the original raw data is important to repeat an analysis. This let us conclude that **there is a lack of knowledge in the data literacy fundamentals where data can be transformed from user-friendly data curation formats to analysis-friendly formats in multiple pipeline steps.**

1.2 Validate learner self-assessment survey. The survey was designed with the questions in duplicate for internal validity, i.e., each construct was asked in 2 separate questions. These questions can be validated using Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA) results. The EFA results can also be used to simplify the survey to a representative set of questions. These questions can be used in future surveys to gain the same amount of information about learners without asking the entire battery of questions in our self-assessment survey. EFA was also used to simplify he survey down to 3 question, one for each of the knowledge domains.

Figure 8 shows how half of the questions (8 out of 16) account for more than 90% of the variance in PCA, with the first component accounting for 46% and 3 components accounting for 68%. Since our survey was designed using 3 main constructs: programming, statistics, and data knowledge, these results show that the survey was well designed and has internal consistency among the respondents. We also conducted an exploratory factor analysis on our preliminary data, and used 3 factors, one for each latent variable. The question loadings also followed the constructs in the survey, suggesting the validity of the survey. We would need to expand the survey to more participants to show its external validity. Given the preliminary data, we propose to perform more validation checks by increasing the sample size. We will be able to calculate the Cohen's kappa coefficient to measure inter-rater reliability and the larger sample size will improve the external validity of the survey.

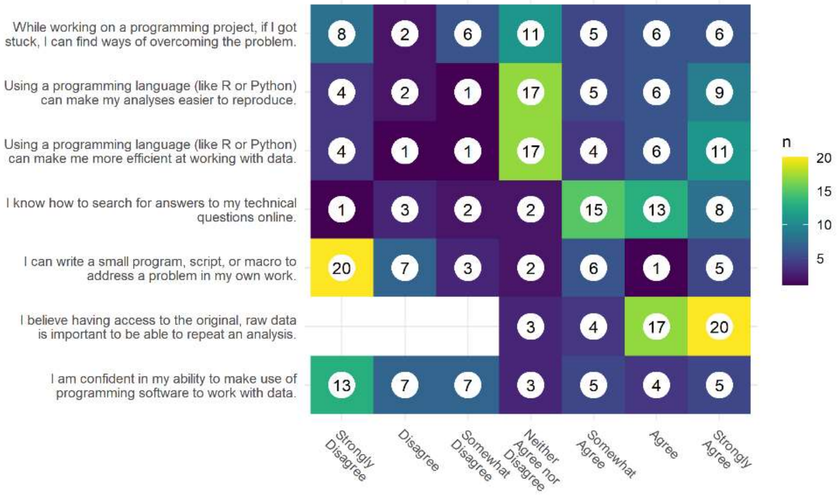

Figure 7. Summary Likert responses. Each respondent was asked about their agreement towards each statement. Preliminary results indicate that each of the personas will have a different set of responses to this Likert table.

1.3 Personas will encompass a student's prior knowledge using survey data. General background, perception of needs, and special considerations will be added to make each learner persona a complete character. We used hierarchical clustering with Euclidean distance and Ward's method on our preliminary data to create the learner personas (Figure 3).

This approach is validated from preliminary results showing the 4 learner personas from our data. We combined these groupings back with the survey occupation demographics to create the relevant prior knowledge portion of learner personas. Our preliminary data returns 4 personas that map on to the different stages of the Dreyfus model of skill acquisition: clinicians (novice), academics (intermediate), students (intermediate), and programmers (experts).

Anticipated results and their impact. Preliminary results indicate that we can validate the learner self-assessment survey and use the survey results to create learner personas. This gives us an overview of the audience we would potentially teach in the medical and biomedical sciences. Since the final step of persona creation combines the demographic information, the base survey questions can be used across other domains, not just the one we are studying. This potentially gives us a tool to accurately gauge data science learners to better create learning materials for their needs.

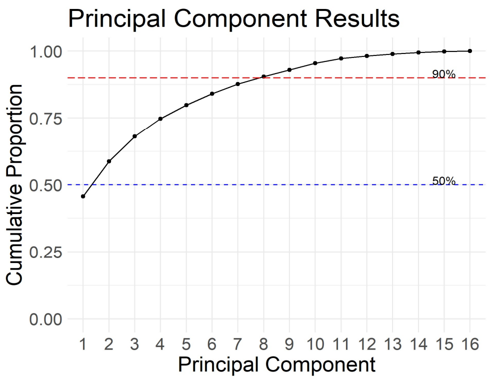

Figure 8. Learner self-assessment validation. Cumulative proportion of explained variance using PCA from the learner self-assessment survey. The survey had questions asked in duplicate for internal consistency. The PCA results show that half the questions (8) account for more than 90% of the explained variance and 1 principal component accounts for 46%. These results suggest that the survey is valid.

Potential pitfalls, alternative approaches, and future directions. The major route of data collection in this Aim is based on survey responses. This inherently means we will have reporting and response bias. The demographic breakdown in our learner self-assessment was diverse (Figure 6), but our preliminary data was only collected from Virginia Tech students and faculty from biomedically relevant listservs. Future directions would include increasing the survey pool to get a better representation of potential learners. A larger survey pool would also help with the survey validation by potentially surveying a more diverse population and increase our N for the analysis. Our initial survey had 57 respondents, where 51 consented to the study, and 45 responses were used for the analysis. These results were from a convenience sample from listservs at the university.

Aim 2: Create an effective data science for biomedical science curriculum based on best education and pedagogy practices. Creating materials that are community-oriented, open, maintained, accessible, follows best pedagogical practices, and domain-specific is a key component to creating authentic tasks to aid in learner motivation. Using the feedback from learners to assess the learning materials, we can create more relevant learning modules, and the surveys and feedback system can be adapted to create other domain-specific learning materials.

Working hypothesis: We hypothesize that a data science curriculum focused on data literacy principles from working with spreadsheet data will be the most relevant to our learners. We also hypothesize that the learning objectives we create will give the learner confidence in performing their own data analysis after going through the materials. By catering to the learner's needs, and teaching the data literacy fundamentals, learners will be more motivated to continue learning on their own. To test this hypothesis, we will create a set of pre-workshop, post-workshop, and long-term workshop surveys. Since this is an observational study, we will use the learner's confidence on their ability to accomplish a task as a proxy for meeting learning objectives. There will be a set of self-assessment and learning objective tasks that will be asked across all surveys to measure differences in response longitudinally.

Preliminary Data for Aim 2. A data science curriculum based on learner personas was created and used to teach a set of workshops. Preliminary data collected before and after the workshop compare a learner's confidence of meeting learning objectives and show that the learners are more confident in their skills and learning objectives (Figure 10). In both figures, the number of responses from the pre-workshop results was subtracted from the number of results in the post-workshop results, giving us the net change in people after the workshop. A blue positive number indicates the shift in the number of participants at the end of the workshop, compared to the red negative numbers pre-workshop.

These results came from 3 workshops, totaling of 72 registrants (Oct: 27, Feb: 45). From this we were able to teach 39 workshop attendees (Oct: 20, 11; Feb: 19, 16) and collected 25 pre-workshop responses, and 12 post workshop responses. We estimate we would need to run at least 12 more workshops to get enough respondents for the post-workshop responses. For the long-term survey (6 months), we may need to provide a financial incentive for participants to get enough responses.

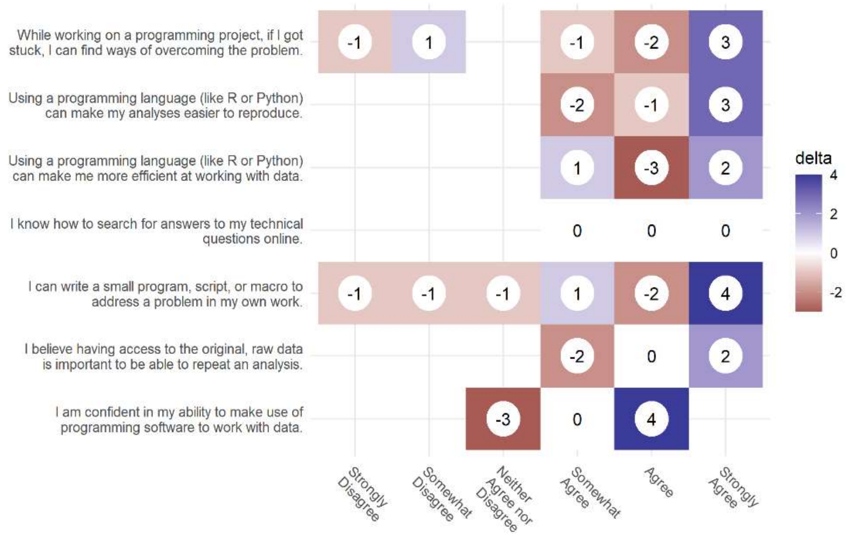

Figure 9. Summary Likert responses pre-post changes. Changes in the Likert table responses from the pre-workshop and post-workshop survey. The delta represents the net differences of responses in the post-workshop survey.

2.1 Learning objectives focused on core data literacy principles in the data science pipeline will be used for each lesson module. In this sub aim, we will create a data science curriculum that ties together data literacy and data management pipelines with the skills used in data science. Our preliminary learner personas guided us to use spreadsheet programs (e.g., Excel, LibreOffice, Google Sheets, etc.) as the first lesson module to orient the learners and use "tidy data principles" as the underlying theme to transform data. Preliminary data and a survey of available lesson materials show that there is a gap in available teaching materials that link data literacy concepts of data management and processing with other steps in the data science process. The initial modules will cover spreadsheets and data literacy basics, loading data into a programming language, tidy data principles, plotting, and logistic regression. Additional modules will cater to the feedback survey from the learners.

2.2 Lesson content follow best educational and pedagogical best practices. The lesson materials created will follow the best educational and pedagogical best practices for learning programming. This includes: (1) using a backwards-design approach to formulate the learning objectives from formative assessment questions (Wilson 2019); (2) blocks of content that fit roughly 50-minute segments to allow for breaks and formative assessments (Farrell and Carey 2018). The materials are also created to work in the context of classroom "periods"; (3) live-coded workshops where the instructor has learners follow along as the concepts are taught and discussed (Farrell and Carey 2018); (4) code of conduct for learners to feel safe in the learning environment (Ambrose et al. 2010; Wilson 2019); (5) live captioning during the workshop event and descriptive alternative text for each figure and image in the online materials to work with screen readers for accessibility. By creating materials that assume no prior knowledge of programming and use relevant examples, students are likely to be more engaged with learning the materials (Farrell and Carey 2018; Ambrose et al. 2010; Wilson 2019).

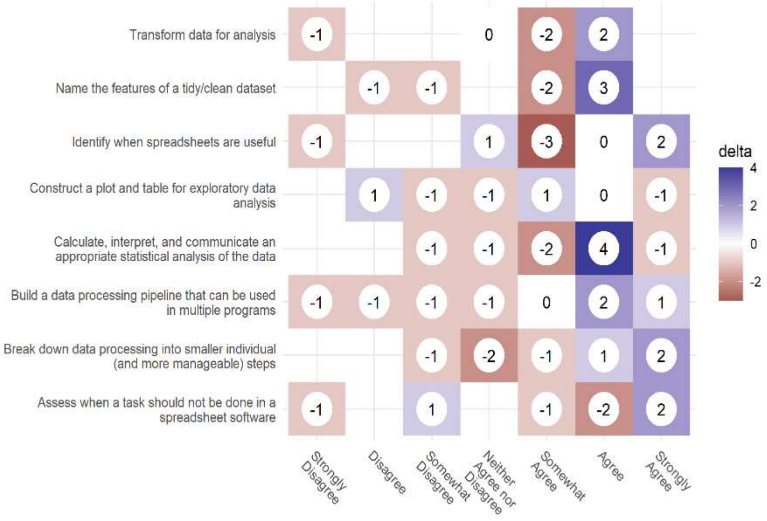

Figure 10. Learning objective pre-post changes. Changes in the learner's confidence in the workshop's learning objectives from the pre-workshop and post-workshop surveys. The delta represents the net differences of responses in the post-workshop survey.

2.3 Assess the effectiveness of learning materials. In this sub aim, we will assess how effective the lesson materials and its learning objectives are by comparing results from a pre-workshop and post-workshop survey. Survey participants will have a unique identifier that can be used to track individual differences and be aggregated to look at the overall effect changes before and after the workshop. Preliminary data shows that learner's confidence in various tasks and learning objectives do improve after the workshop. A long-term survey will be sent out to workshop participants to see retention of learning objectives, if learners found the workshop useful, and if learners have continued to learn and work on their own projects.

Anticipated results and their impact. We anticipate that a lesson curriculum that incorporates existing tools and prior knowledge of spreadsheets will help learners fill in gaps of their data literacy mental model when working with data in data science projects. By focusing on data literacy concepts, we are building a curriculum that promotes FAIR principles. This aim will create a tested learning curriculum that can be adapted into many teachings and learning formats. The book can be done as supplemental reading or as self-paced reading, the slides provide major points that can be used in a lecture or presentation, and recordings are provided to learners either as a reference or material for new learners who cannot attend a live workshop setting. These materials lay the groundwork for a community-oriented, open, accessible, and pedagogically sound curriculum that can be used to enhance the data science and research workforce in the biomedical sciences and adapted to other domains.

Potential pitfalls, alternative approaches, and future directions. Our preliminary data shows more reporting bias than our learner self-assessment survey. Most of the respondents from the workshop surveys are students, and not from the other occupation groups. This problem can be remedied by conducting more workshops to collect more data which may offset the bias. Our surveys mainly measure the learner's confidence towards a learning objective as a proxy for a summative assessment. These results are self-reported and may show response bias.

Aim 3: Assess the effectiveness of formative assessments in learning objectives. Formative assessments "forms" the teaching in real time by informing the instructor what concepts learners are getting wrong (Wilson 2019). At the end of a lesson, a summative assessment can be used to assess learners about all of the individual concepts integrated together. Ambrose et al. (2010) tells us that "goal-directed practice coupled with targeted feedback are critical to learning." Examining if automatic grading systems can be combined with informative feedback can lead to better learning outcomes can identify techniques to improve the democratization of data science education.

Working hypothesis: We hypothesize that formative assessments with targeted and informative feedback about incorrect solutions, will allow learners to complete formative and summative assessment questions with a higher rate of success. We expect that guiding learners with Parson's problems and Faded in formative feedback exercises will help them solve summative feedback questions faster.

Preliminary Data for Aim 3. Our hypothesis is based on computer science education literature that uses different question types for formative assessment questions to aid in learning content. These question types (Faded examples and Parson's problems) are used in lieu of a blank box where the learners write code from scratch because it lowers the cognitive load of the learners and allows them to focus on the key aspect of the coding exercise, instead of wrestling with the syntax of the code.

3.1 Implement an experiment for conducting formative and summative assessment question types. The *shinysurveys* R package provides the framework needed to create and administer an experimental study that can be used to collect response data from user submitted code. It leverages the *learnr* R package that allows instructors to create lesson materials with an input field that can execute code. The *gradethis* library can be used to check R and Python code for the correct result to provide feedback to the student. *gradethis* also can check the syntax of the code itself to point to an exact part of the code that is incorrect, instead of giving a programming error or non-meaningful "incorrect" message. *shinysurveys* can be used in conjunction with the tools and techniques from "Data Science in a Box" to collect the responses from the student for analysis using the *learnrhash* library. These responses are "hashed" such that all the data that is encoded is resented in plain text, but indecipherable to the user.

3.2 Assess the effectiveness of targeted feedback in auto-grading systems used in formative and summative feedback. In this aim, we will show an improvement in the success rate of assessment questions when targeted feedback about the incorrect solution is given by the learner. We recognize the importance of feedback in the learning process (Ambrose et al. 2010), but it is not possible to give real-time feedback during many assessment questions, especially when teaching at scale. We hope to take the results from our implementation of *shinysurveys* and "Data Science in a Box" to collect learner assessment performance and compare the differences between learners who are given several types of assessment questions from those who are simply given an empty box to type code with and without informative feedback from the auto-grader. We currently have a population of students in our own lab, DataBridge, with new data science learners who can be used to take entire workshops and/or short modules where the results from the formative and summative assessments can be recorded.

Anticipated results and their impact. We predict an improvement in speed and correct responses in students' final summative assessment when they are given who are given question types rather than an empty text box in the formative assessment. While these question types are used in computer science education literature, these techniques have not been studied yet to show whether adding the cognitive load of completing a data related task helps with learning the materials. We anticipate that these results will give future educators the types of questions that can be used for formative assessments when teaching.

Potential pitfalls, alternative approaches, and future directions. The study will aim to only teach a single portion of the overall lesson materials, it is possible that the amount of information used for this aim will be either too simple or too complex for participants given the time constraints. If that is the case, we may resort to only looking at the amount of time to complete a solution, rather than comparing if the distinct groups are answering the question correctly. An additional set of workshops are also going to be planned to include the data collection portion of this aim. Since the workshop is going to be more time intensive, there will be a bias in the participants who sign up to attend the study. This aim will provide the basis of incorporating a formative and summative assessment system that can be used in a live teaching environment so the instructor can get feedback about topics and concepts that the learners are grasping. These results will be combined with the demographic information with the learners and can be combined with our persona results to curate a better learning curriculum for our learners.

Project Timeline	**Months 1-3**	**Months 4-6**	**Months 7-9**	**Months 10-12**	**Months 13-15**	**Months 16-18**
Aim 1 (Identify personas)	1.1,1.2, 1.3	1.1,1.2, 1.3				
Aim 2 (Assess learning materials)		2.1, 2.2, 2.3	2.1, 2.2, 2.3	2.1, 2.2, 2.3	2.1, 2.2, 2.3	2.3
Aim 3 (Formative assessment efficacy)			3.1	3.2	3.2	3.2

REFERENCES

American Medical Association. (2021). *Accelerating change in medical education*. https://www.ama-assn.org/education/accelerating-change-medical-education.

Ambrose, S. A., Bridges, M.W., DiPietro, M., Lovett, M.C, & Norman, M.K. (2010). *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons.

Dolgopolovas, V., & Dagienė, V. (2021). Computational thinking: Enhancing STEAM and engineering education, from theory to practice. *Computer Applications in Engineering Education,* 29(1): 5–11. https://doi.org/10.1002/cae.22382

Farrell, K.J., & Carey, C. C. (2018). Power, pitfalls, and potential for integrating computational literacy into undergraduate ecology courses. *Ecology and Evolution,* 8(16): 7744–7751. https://doi.org/10.1002/ece3.4363

Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. (2010). Clinical data as the basic staple of the learning health system. In *Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary* [eBook edition]. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK54306/

Jordan, K.L. (2016, October 20). Data carpentry assessment report: Analysis of post-workshop survey results. *Zenodo*. https://doi.org/10.5281/zenodo.165858

Jordan, K.L., Marwick, B., Duckles, J., Zimmerman, N., & Becker, E. (2017a, July 1). Analysis of software carpentry's post-workshop surveys. *Zenodo*. https://doi.org/10.5281/zenodo.1043533

Jordan, K.L., Marwick, B., Weaver, B., Zimmerman, N., Williams, J., Teal, T., Becker, E., Duckles, J., Duckles, B., & Wickes, E. (2017b, October 31). Analysis of the carpentries' long-term feedback survey. *Zenodo*. https://doi.org/10.5281/zenodo.1039944

Jordan, K.L. (2018, July 17). Analysis of the carpentries long-term impact survey. *Zenodo*. https://doi.org/10.5281/zenodo.1402200

Jordan, K.L., Michonneau, F., & Weaver, B. (2018, July 17). Analysis of software and data carpentry's pre- and post-workshop surveys. *Zenodo*. https://doi.org/10.5281/zenodo.1325464

Jordan, Kari L., François Michonneau. (2020, March 26). "Analysis of the Carpentries Long-Term Surveys (April 2020)." Zenodo. https://doi.org/10.5281/zenodo.3728205.

Kross, Sean, Roger D. Peng, Brian S. Caffo, Ira Gooding, and Jeffrey T. Leek. 2020. "The Democratization of Data Science Education." *The American Statistician* 74 (1): 1–7. https://doi.org/10.1080/00031305.2019.1668849.

Milo, S. (2005). Information literacy, statistical literacy, data literacy. *IASSIST Quarterly* 28(2-3): 6–6.

Moriarty, A. (2020, May 26). Does hospital EHR adoption actually improve data sharing? *Definitive Healthcare*. https://blog.definitivehc.com/hospital-ehr-adoption.

Ogier, A., Brown, A.M., Petters, J., Hilal, A. & Porter, N. (2018). Enhancing collaboration across the research ecosystem: Using libraries as hubs for discipline-specific data experts. *Proceedings of the Practice and Experience on Advanced Research Computing, 60,* 1-6. https://doi.org/10.1145/3219104.3219126.

Office for Civil Rights (OCR). (2017, June 16). HITECH Act enforcement interim final rule. HHS.gov. https://www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/index.html

Office of the National Coordinator for Health Information Technology (ONC). (2021). *Health IT legislation*. https://www.healthit.gov/topic/laws-regulation-and-policy/health-it-legislation

Pruitt, J., & Adlin, T. (2006). *The persona lifecycle: Keeping people in mind throughout product design*. Morgan Kaufmann.

Song, I., & Zhu, Y. (2016, August). Big data and data science: What should we teach? *Expert Systems,* 33(4): 364–373. https://doi.org/10.1111/exsy.12130

Stewart, C. (2020, September 24). Healthcare data volume globally 2020 forecast. *Statista*. https://www.statista.com/statistics/1037970/global-healthcare-data-volume/

U.S. Dept. of Health and Human Services. (2017). *HITECH act summary*. https://www.hipaasurvivalguide.com/hitech-act-summary.php

Wickham, H. (2014). Tidy data. *Journal of Statistical Software* 59 (10), 1–23. https://doi.org/10.18637/jss.v059.i10

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. [eBook edition]. O'Reilly Media. https://r4ds.had.co.nz/

Wilson, G. (2019). *Teaching tech together: How to make your lessons work and build a teaching community around them*. CRC Press.

Zagallo, P., McCourt, J., Idsardi, R., Smith, M.K., Urban-Lurain, M., Andrews, T.C., Haudek, K., Knight, J.K., Merrill, J., Nehm, R., Prevost, L.B., Lemons, P.P. (2019, November 22). Through the eyes of faculty: Using personas as a tool for learner-centered professional development. *CBE—Life Sciences Education,* 18(4): ar62, 1-21. https://doi.org/10.1187/cbe.19-06-0114