Received: 31 January 2022 | Revised: 5 July 2022 | Accepted: 25 July 2022

DOI: 10.1002/bmb.21663

ARTICLE

Biochemistry and Molecular Biology Education

WILEY

Pedagogy of teaching with large datasets: Designing and implementing effective data-based activities

Catherine M. O'Reilly[1] | **Tanya Josek**[2] | **Rebekka D. Darner**[2,3] | **Sarah K. Fortner**[4]

[1]Department of Geography, Geology, and the Environment, Illinois State University, Normal, Illinois, USA

[2]School of Biological Sciences, Illinois State University, Normal, Illinois, USA

[3]Center for Mathematics, Science and Technology, Illinois State University, Normal, Illinois, USA

[4]Science Education Resource Center, Carleton College, Northfield, Minnesota, USA

Correspondence
Catherine M. O'Reilly, Department of Geography, Geology, and the Environment, Illinois State University, Normal, IL, USA.
Email: cmoreil@ilstu.edu

Funding information
National Science Foundation, Grant/Award Numbers: IUSE-1821564, IUSE-1821567

Abstract

Integrating the use of large datasets into our teaching provides critical and unique opportunities to build students' skills and conceptual knowledge. Here, we discuss the core components needed to develop effective activities based on large datasets, which align with the 5E learning cycle. Data-based activities should be structured around a relevant question, use authentic publicly accessible data, be scaffolded to include choice, and involve discussion of the results. It is important that the software that is used to manipulate, analyze and/or visualize the data is accessible for students. There are a range of strategies to reduce the barriers of working with large datasets through pre-organizing and pre-scripting code for analyses, using online cloud-based versions of software, and reducing opportunities for error in syntax. Resources exist for learning open-source software (e.g., Data Carpentry) as well as for support and professional development in teaching with large datasets (Project EDDIE).

KEYWORDS

big data, professional development, RStudio

1 | VALUE OF WORKING WITH LARGE DATASETS

Teaching with large datasets allows us to build students' skills and conceptual knowledge in multiple ways. Working with large datasets allows students to work with datasets well beyond the scope of what they could collect in a classroom and necessitates the development of quantitative skills. Students can learn practices associated with the manipulation and visualization of large datasets while simultaneously building their conceptual understanding by interpreting their results within a disciplinary context. In particular, working with publicly accessible data democratizes science and provides opportunities for students to improve their quantitative reasoning (the ability to create quantitative evidence and use it to support a relevant argument). Quantitative reasoning is a critical skill needed to empower citizens, promote healthy social discourse, support solid democracies, and prepare students for almost every career path in the 21st century.[1,2]

Using large datasets mirrors what we do as scientists and allows students to develop competencies with understanding data. It provides an opportunity for students to

This article reports a session from the virtual international 2021 IUBMB/ASBMB workshop, "Teaching Science on Big Data."

learn how and why scientific data are reliable and valid and how to reason with supporting evidence. Doing so also gives students the opportunity to develop the skills necessary to evaluate the quality of a dataset, which is a prerequisite to developing accurate beliefs about the nature and validity of scientific knowledge. These epistemic beliefs about science, such as students' beliefs about the sources of scientific knowledge, how scientific knowledge is verified, and the role of uncertainty in science, directly influence their ability to reason with evidence.[3] Without healthy epistemic beliefs about science, students are unlikely to be able to distinguish between the validity of scientific claims that are based on empirical evidence and those based on anecdotal evidence or no evidence at all.

In addition to skill development, working with large authentic datasets provides opportunities for application and understanding of scientific concepts. Engaging with large datasets, which are messy, compels students' decision-making on a variety of analytical judgments that otherwise do not occur. For example, when they encounter outliers and have to make decisions about whether to include them in an analysis, students are compelled to think about variation caused by natural change in the phenomenon under study versus measurement error. Additionally, when large datasets are used, no matter what conclusions are drawn, they will likely be correct and help correct any misconceptions about the underlying concepts. Unexpected results can help students identify and resolve their misconceptions. Faced with cognitive dissonance and a need to answer a relevant question, students revisit their schema and seek explanations for why the data do not match their preconception, which well-positions them to discover why their previous schema was wrong. Thus, students have to grapple with understanding their results; it is not an option to say "more data is needed." Extensive evidence has demonstrated that learning is most effective when situated in an authentic context,[4] because students are more motivated when they can imagine how they will use what they are learning in the future.

Teaching with large datasets also provides opportunities to improve quantitative reasoning in students. Using quantitative reasoning requires that individuals think critically and apply basic quantitative and statistical thinking skills to interpret data, draw conclusions, and solve problems within a disciplinary context.[5] This includes quantitative skills such as manipulating numbers and knowing what analyses to do (also called numeracy). It requires critical thinking—results should be interpreted and incorporated into an argument based on the data to address a question (also considered quantitative literacy). Finally, the question should have some meaning and be rooted in a disciplinary context.[6] In actuality, this is "science in action"—what we do as scientists every time we present at a conference or write a manuscript. Thus, quantitative reasoning involves quantitative skills, quantitative literacy, and disciplinary context, and it is critical to integrate this full suite of activities into how we teach science. Teaching quantitative skills without context devalues the learning experience, making it less likely that students will remember that skill and be able to apply it.

2 | DESIGNING EFFECTIVE ACTIVITIES WITH LARGE DATASETS TO PROMOTE QUANTITATIVE REASONING

In addition to using publicly accessible data (discussed above), effective data-based activities should (a) be structured around a relevant question, (b) be scaffolded to include choice, and (c) involve discussion of the results. These components of an activity follow the highly effective 5E model for teaching science (Figure 1).[7,8] The 5E learning cycle first *engages* students, which is most easily done with an appealing question and a relatively simple exploration of the data that makes it seem accessible. The subsequent steps are to *explore*, *explain*, and *expand* (Figure 1).[7] Implementing these steps effectively will require that students have some independence and choice as part of the activity. The last stage of the 5E is *evaluation*, which can be formative assessment for both the student and the instructor, as well as summative assessment at the end of the activity. Ultimately, to demonstrate quantitative reasoning, the student needs to be

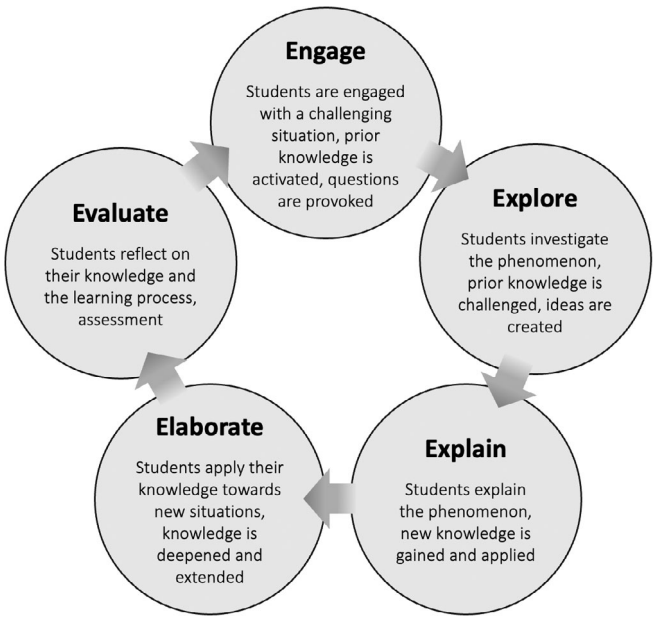

FIGURE 1 The 5E learning cycle showing the progression of activities, starting with Engage.[7,8] This model can serve as a guide for the design and implementation of data-based activities

able to address the overarching question using quantitative information that they collected to support their answer. Activities should also be designed around the instructors specific learning goals, and in practice, the process of designing a data-based activity can be highly iterative.

2.1 | An overarching, relevant question provides a guide

An overarching question acts to drive and guide the activity. Question-driven investigation mirrors what we do as scientists and the scientific process. When working with large datasets, it can be easy to get caught up in the quantitative skills and explaining software, so an overarching question helps data-based activities retain a connection to the disciplinary content and makes the activity more meaningful for students (they know *why* they are doing something). Faced with a large complex dataset, it can be easy to get distracted by details in the data, and having an overarching question can keep the activity focused and grounded. Ideally, questions should be framed in a way that is relevant to disciplinary content. For example, "How do we use data to understand climate change?" is a lot less appealing than "How do we know that humans have contributed to current climate change?" Thus, the overarching question also acts to engage the students, the first step in the 5E learning cycle (Figure 1).[7] Overall, this inquiry-based learning approach promotes engagement, curiosity, and experimentation, where students are empowered to explore subjects by asking questions and finding or creating solutions (Figure 2).

2.2 | Scaffolding activities increases student independence

Providing support through scaffolding is a key component in teaching new tasks with multiple steps and growing complexity. Scaffolding involves breaking up an activity into distinct components that build on each other. Scaffolding a data-based activity allows students to learn techniques and develop self-confidence before independent exploration. Early steps in an activity based on a large dataset can be overwhelming for students, as they are trying to understand the data and learning new analytical techniques, all while expecting to be able to situate this within the context of disciplinary content that probably still seems abstract. Asking students to generate a hypothesis about what they expect to find before they undertake data-based tasks can help students know what to look for, as well as helping instructors identify misconceptions early on. Keeping initial steps in an activity simple and

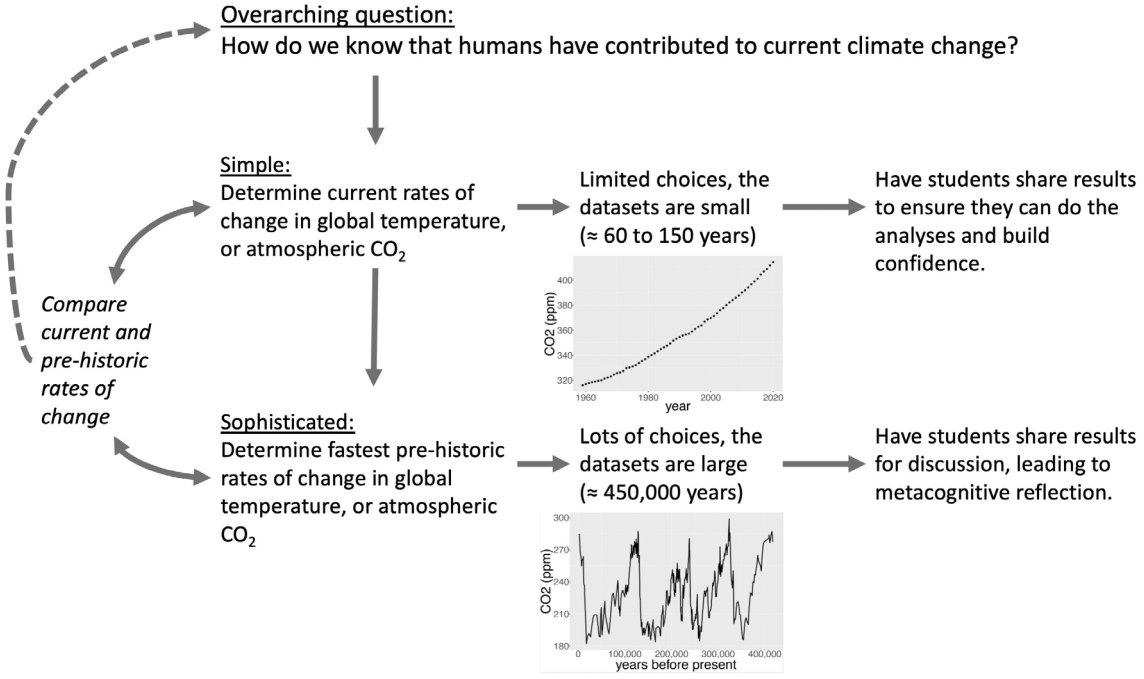

FIGURE 2 Question-driven activities are most likely to be successful. This conceptual outline of a Project EDDIE module focuses on climate change, one of the most widely used modules. The overarching question engages students and provides some goalposts. Students learn the quantitative skills first in a simple exercise with limited data. Sharing those results is formative feedback for students and instructors. This gives students confidence that they can conduct the same analysis by choosing part of a larger dataset. Even though students will each get slightly different results depending upon their choices, their overall answer to the overarching question will be the same

allowing time for students to confirm that they are successful will build their self-confidence. Ideally, initial steps in an activity would introduce the students to the data, maybe just through a simple visualization, and would also act to engage them. Secondary steps in an activity would progressively introduce more advanced quantitative skills and sophistocated interpretation (Figure 2).

Independent exploration is critical for students to reinforce newly acquired quantitative skills and to test their conceptual understanding. It is this, the most sophisticated step of the scaffolding, that is often the most engaging for students. When students take control of how to explore and make sense of data, they are afforded more opportunities to build their self-efficacy (the belief that one experiences when they feel capable of accomplishing their goals[9]) as scientists and in their ability to reason from data.[10] Given a relevant question and a large dataset, students gain more from the experience when they have ownership, rather than when it is mimicking authenticity ("cookbook") and they are expected to determine a "right" answer.[4] It is important to note that this independent exploration should build on the quantitative skills students already developed earlier in the activity—for example, they are doing a similar analysis but are choosing the data to work with (Figure 2).

However, allowing students choices can be intimidating for instructors. Because students will be independently choosing some aspect of the data or analyses and their results will not be known in advance, instructors also have to be comfortable thinking and teaching in the moment. Carefully designed activities ensure that students will be working within known constraints, facilitating the instructors ability to anticipate the types of results that students might get. An instructor's more advanced knowledge and prior practice of working with these concepts and data make it highly likely that an instructor will be able to work though understanding an unexpected result; in fact talking through this process out loud is an excellent way for students to see how scientists work. It is also ok not to know the answers when something unexpected happens—this is part of doing science as well, taking time to figure out what is happening when an unusual result occurs; instructors should feel ok saying that they will need more time to think about this and that they will follow up in a future class session.

2.3 | Discussion facilitates the development of scientific argumentation and metacognitive reflection

Finally, communication is a key part of metacognitive reflection. The best reflective assignments respond to an authentic problem or a disagreement that needs to be resolved. Having students share their answers as they work through an activity is one way to promote metacognition. With the initial, simpler activities, students should probably get similar answers, and comparing their answer to their peers' builds self-confidence in their quantitative skills. As an instructor, comparing results for these simple activities also provides a way to monitor progress, check that students are capable of conducting the desired analyses, and catch any misunderstandings. With the more sophisticated activities, comparing answers allows students to see that while there is no "right" answer there is a consistent pattern across everyone's results. Students learn that the choice of data does influence the outcome to some extent, and the opportunity to explain and defend their choices during discussion allows them to reflect on their decision-making process. Because students are unlikely to all have the exact same answer for the last part of the activity, this creates an opportunity for discussion that will encourage students to reflect on their answer and their process for generating it.

3 | PRACTICAL ISSUES WITH SOFTWARE IN THE CLASSROOM

Even with the best learning intentions, there are still hurdles that need to be overcome especially when using data-based activities in the classroom. In order to analyze data, specific types of software need to be used and while these software exist, in order to conduct these data-based activities instructors are faced with difficult decisions. Instructors can be limited on what types of software they have available to them, how much time they have to learn new software, what software their students have access to at school and at home, and the useability of different software.

Software programs can be incredibly inaccessible to individuals. Inaccessibility can come in a variety of forms. One of the main ways software can be inaccessible would be due to the fact that many popular statistical analysis programs, such as SPSS or SAS have moved to cloud-based subscription programs which cost users ($1100–$3000 per year). Statistical programs such as Python (www.python.org), R, RStudio (https://www.rstudio.com), and R-Studio Cloud (https://rstudio.cloud; a cloud-based online version of R) are exceptions to this, as they are open-source and freely available. There are RAM limitations to using R-Studio cloud for free, but they have multiple tiered options to fit an individual or classroom needs that minimize cost ($60–90 per year). Spreadsheet programs (Excel, GoogleSheets, etc.) can also be used as a more affordable option for data analysis, but they are limited in their statistical analysis capabilities.

Another way software can be inaccessible is the fact that there are few software that allow individuals with accessibilities needs (e.g., screen readers, voice-to-text, alternative controls, etc.) to use their programs. R is one of the few software that offers packages and programs that allows for individuals with varying accessibility needs to successfully use their programs.[11]

Accessibility can also be limited by what resources the user has available to them. While many individuals have access to personal computers and the internet, this is not universally true. Some individuals are limited to what they have at their public library, which goes back to software access. Library computers are not likely to have expensive statistical software available and users cannot download software onto these computers. In order to use RStudio Cloud, the user only need access the internet. Having a computer can be beneficial, but it does not need to be a personal computer because RStudio Cloud is not a desktop software. This allows the users to access their data and analyses from any computer with internet access. It also facilitates getting started, because the instructor can avoid issues associated with getting students to install software across their different platforms.

Another limiting factor that applies to any software is the time to learn it. All software used for data-based activities in the classroom will take time for both the instructor and students to learn. Instructors will often have to put in extra time to not only become comfortable with the software, but to learn it enough that they can help their students when they are stuck, have error messages or are not able to analyze the data. In some cases, it is important that instructors are familiar with both spreadsheet software and data analyses software to complete an activity. In terms of learning software, there are a large number of resources online with tutorials on how to use the software (e.g., Data Carpentry at datacarpentry.org). Open source software, like R, can allow for even more resources to be available because code written for the software is able to be openly shared and modified. This means that individuals can write specialty code to modify R software which can be openly shared and freely downloaded and used at the users discretion, although this can create another challenge for instructors in terms of being up-to-date on the most widely used packages.

There are a range of practical strategies to facilitate how students engage with R software, particularly for those who are being exposed to coding for the first time (Box 1).[12] The simplest approach is to use RStudio Cloud which avoids issues of installation across different platforms and provides instructor access to student work that can facilitate asynchronous troubleshooting. Through a single shared link from an instructor, students can be looking at pre-organized projects opposed to having students navigate through multiple links and pages. Providing pre-written code makes it easier for students to focus on the analyses and results without getting frustrated by trying to understand the language and how to code. Pre-written code also allows you to teach at a variety of difficulty levels, for example, if you want to have an easier lesson, you can have all of the code available and guide students through how to change their code for their data. If you want a more challenging project you can give students a base code example and then have them decide how the code needs to be changed based on the previous code given. If students are provided with a pre-organized R Project, then all that is needed is to hit "Run" to move through the code. The ease at which visualizations appear engages students and reduces their anxiety.

It is relatively simple to provide pre-organized R Projects via Github, or through sharing a direct link to a Workspace in RStudio Cloud where everything is already set up. Students will need an orientation to the space, so that they know how to get started by opening the scripts folder and the script file, and to learn what to do to run code. Although both instructors and students might be nervous about using code-based software in introductory courses, we have found that most issues involve (a) packages that were not installed or (b) incorrect syntax. One additional benefit of R platforms is that the software does have an intuitive "warning" system built in so that if any of the code has errors such as missing quotations or incorrect punctuation, red lines or dot will display on the line of code to indicate that the code cannot run. Additionally, when code is run that does not work, the error message provided by the R software (regardless if desktop or cloud) will provide detailed information about the error which makes it easier for both the

BOX 1 Practical strategies for teaching with RStudio

1. Pre-organize projects for easier student navigation.
2. Provide code and tips for code modifications that students can practice with.
3. Use online R-platform RStudio Cloud rather than relying on student installation of software.
4. Provide instructions and support for student troubleshooting.
5. Consider assigning students who work through problems quickly as peer mentors.

instructor and student to determine why the code failed to run. Even though we may be providing pre-scripted code, it is important to still provide students with opportunities for choice in more sophisticated sections of the activity. As part of the scaffolding, later sections of the activity can include place-holder text within the code indicated in a way such that students know what text they should change. Instructors can use commenting for directions, and ideally the code would be similar to some that had already been done earlier in the activity for comparison.

If possible, we recommend having a teaching assistant or second instructor present who can help students when they encounter a problem with their code. Having a second instructor is a strategy strongly recommended by Data Carpentry, an organization that develops and teaches workshops on data skills (datacarpentry.org). You may also be able to have students who finish first then act as teaching assistants to help other students still working on the activity. For virtual teaching, a separate breakout room can be set up where students can join when they need help. It is also recommended that you tell your students to please join the virtual class and run RStudio Cloud (or other software) on the same computer so that students can share their screen. Many students prefer to join virtual classrooms on their phones and complete the activity on their computer, but this limits their ability to share their screen. It is important for students to be able to share their screen it make problem solving easier for both that student and for the class as many times when one student has problem, a few others may be experiencing similar issues.

Although software can act as barriers, they are not the only factors that cause instructors to be wary of data-based activities. Because of the lack of data-based activities with fully developed lesson plans, some instructors may be less likely to use these activities because they are not able to provide a set of consistent instruction that cover the diverse set of content in their course. Finally, data-based activities do take time, and some instructors would prefer to cover more content as opposed to allowing students to form the deeper concept understanding which these data-based activities promote. However, greater learning is possible with a focus on practicing core quantitative reasoning competencies embedded in EDDIE modules such as: interpretation, representation, calculation, analyses, synthesis, assumptions, and communication.[13] Part of increasing instructor buy-in may therefore include providing orientation in how students learn inquiry and quantitative reasoning and support for trying out new teaching approaches.

4 | COMMUNITY-BUILDING THROUGH SHARED RESOURCES AND SUPPORT

There is growing interest in teaching using large datasets and opportunities to get support and training to do so. Project EDDIE (Environmental Data-Driven Inquiry and Education, projecteddie.) is a community of STEM disciplinary and educational researchers that provides support for instructors wanting to teach with large datasets. Modules have been developed by instructors entering in at various levels of comfort teaching with large, openly available environmental datasets. In addition, Project EDDIE provides professional development and community-building activities to support instructors getting started using EDDIE modules and approaches that are new to them. Professional development, including workshops and webinars on using EDDIE materials, appeals to instructor extrinsic motivation and provides opportunities to work through new approaches in supportive environments.[14]

Although Project EDDIE initiated around supporting faculty in creating teaching modules that embed datasets that relate to environmental issues, the ideas guiding the design of the modules also apply to structuring support for faculty that are just getting started teaching with large datasets. These modules are designed with scaffolded parts A, B, and C that match the 5E learning cycle. They are flexible and can be adapted for a range of course levels, and can be completed over the course of 3 h, which can be any combination of several lecture periods, lecture and homework, or a laboratory session. These modules have been assessed and are effective at improving students competence and willingness to work with large datasets.[15] In parallel, faculty development provides practice and exposure to pedagogy, constructing effective activities, and uniting around big data inquiry and quantitative reasoning that instructors with similar goals and topical foci find appealing.

An implicit goal of Project EDDIE is a desire to build both instructor and student comfort engaging in inquiry with openly available data sets. While initially this was about materials development, it now includes workshops, webinars, and instructor mentoring communities. All of these embed opportunities to practice effective pedagogical approaches by using modules or reflecting on elements of their construction, or how to operationalize a module in an instructor's teaching context. Instructors learning how to teach with inquiry value learning the underpinning science, opportunities to practice what they are learning and accountability for applying ideas.[16] Our workshops begin by sharing the evidence-based framing for EDDIE module construction and then engage

participants in the materials. Through this they learn strategies for supporting students in inquiry and quantitative reasoning. They also have the opportunity to reflect and discuss sticking points, or how to customize ideas into their courses. These sharing opportunities highlight the special expertise of each participant's context and begin to shape enthusiasm and accountability for applying EDDIE ideas. They might also help instructors overcome fear of losing content in favor of student-centered approaches because engagement and discussion highlights their value and use.[17]

Creating and curating teaching materials around common community-identified interests (e.g., teaching with open data) encourages those who have engaged in Project EDDIE professional development or learned about EDDIE materials from their colleagues to explore the full breadth of resources on the website that might apply to their courses. Likewise, instructors who learn about Project EDDIE by attending a workshop or webinar are shown where and how to find other materials and opportunities on the website. By design, we have created a feedback between resource sharing and professional development that generates interest and use of materials. Even people who find the resources by searching on their own are connected to other elements of the project by website design, tagging, and search functionality.

5 | CONCLUSION

Using data-based activities in the classroom can be an effective way to engage students in the material while also learning quantitative skills and practicing quantitative reasoning. Well-designed activities that follow the 5E learning cycling and provide scaffolding can provide students with substantive learning experiences, making the trade-off between covering content in lecture compared to spending time working on an activity well worth it. Strategies for using the coding software needed to work with large datasets can reduce frustration for students and instructors, minimizing this barrier to focusing on the results and scientific concepts. We encourage instructors to develop their own collaborative communities to get ideas, feedback and share their successes. Resources exist for learning more about software as well as professional development for teaching with data.

ACKNOWLEDGMENT

Project EDDIE is supported by NSF IUSE-1821567 and IUSE 1821564.

ORCID

Catherine M. O'Reilly https://orcid.org/0000-0001-9685-3697

REFERENCES

1. National Research Council. Report of a workshop of pedagogical aspects of computational thinking. Washington, DC: The National Academies Press; 2011.
2. Partnerships for 21st Century Learning. Communication and Collaboration; 2016. Available from: http://www.p21.org/about-us/p21-framework/261
3. Fang SC, Hsu YS, Lin SS. Conceptualizing socioscientific decision making from a review of research in science education. Int J Sci Math Educ. 2019;17:427–48.
4. Rule AC. The components of authentic learning. J Authent Learn. 2006;3:1–10.
5. Mayes RL, Forrester J, Christus JS, Peterson F, Walker R. Quantitative reasoning learning progression: the matrix, numeracy: advancing education in quantitative. Literacy. 2014;7:1–20.
6. Elrod S. Quantitative reasoning: the next "across the curriculum" movement. Peer Review. 2014;16:4–8.
7. Bybee RW, Taylor JA, Gardner A, Van Scatter P, Carlson Powell J, Westbrook A, et al. BSCS SE instructional model: origins and effectiveness, a report prepared for the Office of Science Education, National Institutes of Health. Colorado Springs, CO: BSCS; 2006.
8. Duran LB, Duran E. The 5E instructional model: a learning cycle approach for inquiry-based science teaching. Sci Educ Rev. 2004;3:49–58.
9. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. In: Baumeister RF, editor. The self in social psychology. New York: Psychology Press; 1999.
10. Geitz G, Joosten-ten Brinke D, Kirschner PA. Changing learning behaviour: self-efficacy and goal orientation in PBL groups in higher education. Int J Educ Res. 2017;75:146–58.
11. Godfrey AJR. Statistical software from a blind person's perspective. R J. 2013;5:73.
12. Auker LA, Barthelmess EL. Teaching R in the undergraduate ecology classroom: approaches, lessons learned, and recommendations. Ecosphere. 2020;11(4):e03060. https://doi.org/10.1002/ecs2.3060
13. Madison BL. How does one design or evaluate a course in quantitative reasoning? Numeracy. 2014;7:3.
14. Bouwma-Gearhart J. Research university STEM faculty members' motivation to engage in teaching professional development: building the choir through an appeal to extrinsic motivation and ego. J Sci Educ Technol. 2012;21:558–70.
15. O'Reilly CM, Gougis RD, Klug JL, Carey CC, Richardson DC, Bader NE, et al. Using large datasets for open-ended inquiry in undergraduate classrooms. Bioscience. 2017;67:1052–61.
16. Jeanpierre B, Oberhauser K, Freeman C. Characteristics of professional development that effect change in secondary science teachers' classroom practices. J Res Sci Teach. 2005;42:668–90.
17. Borda E, Schumacher E, Hanley D, Geary E, Warren S, Ipsen C, et al. Initial implementation of active learning strategies in large, lecture STEM courses: lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. Int J STEM Educ. 2020;7:1–18.

How to cite this article: O'Reilly CM, Josek T, Darner RD, Fortner SK. Pedagogy of teaching with large datasets: Designing and implementing effective data-based activities. Biochem Mol Biol Educ. 2022; 50(5):466–72. https://doi.org/10.1002/bmb.21663