Summary Report on Scaling the Introduction to R for Biologists Workshop by National Center for Genome Analysis Support (NCGAS) to a Massive Open Online Course (MOOC)

Sheri A. Sanders
Thomas G. Doak
Carrie L. Ganote
Bhavya Papudeshi

Indiana University
PTI Technical Report

December 10, 2019

Introduction:

A common request on our user survey is to have a webinar or video tutorials. Another common request is to have R, Python, and other basic programming skills workshops for the community. As a result, we have been working to offer a mix of material delivery while still maintaining our ability to scale to the national audience. This effort has included YouTube videos, online office hours, a full online course, in addition to in-person workshops at different locations in the country, often corresponding to conferences our clientele would already be attending.

Implementation:

Design:

There are several other classes offered at IU and other institutions that offer experience with R in applied fashions, however, there is a unfilled need to start from the absolute basics and focus on the programming aspects of the language – accessing the environment, declaring variables, core datatypes. While python and Linux command line are often taught this way, there appears to be a general lack of this level of instruction, making R seem much less approachable than other languages. We focus on unifying concepts across computing languages, because it makes reading and writing R scripts easier, especially if the participants have prior experience with other languages, or if they do not, it makes it easier in the future to pick up other languages (like Python).

While starting at the basics, this course was designed for flexibility from the start. Further, the course is split up into separate modules, allowing us to modify the course easily. Each "concept" has two parts:

- a lecture section with demos and lots of explanations, metaphores, and ties to biological concepts, and
- a lab section that directly references the learning objectives, lecture material, and an example biological application.

These parts have been designed to allow for either a traditional format, with live lectures and follow-up labs, or a "flipped classroom" in which the students can watch the lecture videos from the live lectures and join us in live or live-over-the-web coding sessions as they work through the labs. The concept modules can also be rearranged, subselected, or used individually to taylor the material to different groups and needs.

Scaling:

With the course designed for flexibility, we were able to scale this course from 30 students to 330 students over two years and four offerings. The scaling followed these stages:

- **Initial offering of in-person**: The course was run locally with a group of 30 students over six days. This included two concept modules, one on core datatypes and one on advanced topics such as loops and functions. The labs and lectures were given on separate days. Pre- and post-course surveys were given to evaluate the format and improve the materal. A third module was developed from separately given material on mapping in GIS and ordination graphing for metagenomics, with similar evaluation.

  Dates: March 27,29; April 3,5 2018 – four part
  August 27,29; September 10,12 2018 – six part
  November 16, 2018 – four part at Texas Christian University

- **Hybrid format**: Feedback from the first stage was used to solidify the material. Three full concepts were finalized and the lecture notes were converted into a full 70-page LaTeX-based textbook. As the initial offering had XXX% oversubscription, we extended the offering to be a

hybrid format. Students took the course either at IUPUI with the main instructor, IUB telecast but with instructors to assist with labs/questions/etc., and fully online with telecast lecture and online office hours/help. The course was also reformated into three days, each with a full module (basics, graphing, advanced) to accommodate travel between campuses. We maintained the same pre- and post-course surveys and recorded all lectures. Lectures were split by section in the textbook, cleaned up by IU's ITCO, and uploaded to IU_PTI YouTube.

Dates: February 11, 18, 25 2019 – six part

- **MOOC format**: With the material fully fleshed out into a textbook and videos recorded, the course was converted into an online course, accessible through IU's Expand Learning Management System. This system overlies Canvas and allows non-IU students to access the material at any time. However, as direct interaction was a clear priority of the participants (via the surveys), we organized an entirely online offering of the course. This was managed through Piazza discussion boards, which is accessed through Expand/Canvas. Instructors were available online for set amounts of time three days a week for two week (a total of six times to correspond to the six lab/lectures and chapters in the textbook). This eliminated the travel burden and allowed us to scale instruction to hundreds of participants without increasing teaching time. The same pre- and post-course evaluations were used.

  Dates: November 4, 6, 8, 11, 13, 15, 2019 – 6 part

In addion to format, platform was also considered in scalability. The course uses XSEDE's Jetstream Cloud system to host publicly available virtual machines with Apache, RStudio Server, and bioconductor pre-installed. These images are regularly maintained by NCGAS staff for security and package updates. By using this platform, we were able to avoid using instructor time to troubleshoot various installations of R/RStudio, the variability of which only increases with scale of course. Additionally, the choice to use RStudio Server allowed us to own the virtual machines used in class on our education allocation, while requiring only a web browser for the students to access the software. New virtual machines could be swapped in if necessary and instructors could log into any machine to troubleshoot. Finally, as the computation needed for the course is not demanding, instructors could pack three students onto each virtual machine, allowing for a much more efficient use of educational allocation SUs.

Persistence:

Instead of building the course directly into Canvas/Expand, EOT piloted a development paradigm in which the course was built into our Knowledge Management System (KMS), which features an API that allows us to pull the material directly into Canvas/Expand or any other learning management software the university may move to. As such, all material has redundant form of access between the videos being offered on YouTube, the textbook being offered as a PDF, and the material being hosted on KMS (Figure 1).

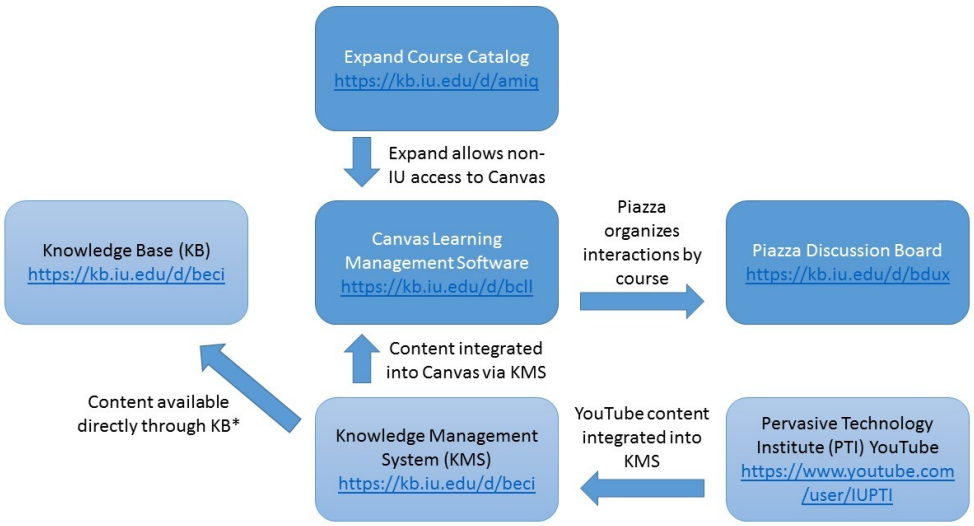

Figure 1: Basic layout of course infrastructure for Expand. The light blue boxes are persistent back-up access modes for the main content platforms (dark blue boxes). For instance, if Canvas/Expand were to be replaced by anything, content could be immediately available on YouTube as well as KB. Once a replacement is implemented, content could be pulled back in via the KMS API. * indicates that while this aspect is possible, it is currently not implemented as directly accessible to public.

Teams Involved:

National Center for Genome Analysis Support (NCGAS) – main developers, managers, and instructors of the course materials

Jetstream Project Management and Research Infrastructure – assisted in setting up virtual machines and allocations for course

Collaboration and Engagement Support (CES) – managed migration of materials to KMS, scheduled rooms and other necessary infrastructure to teach, managed announcements and advertisement, provided branding materials, and managed the registration via IT Training's SuperComputing for Everyone.

IT Communications (ITCO) – Assisted CES in providing branded materials and advertisements, performed all video editting.

Canvas/Expand Team – assisted with migration of materials to Expand/Canvas via KMS

Recruitment:

NCGAS contacted all current and previous clients and all participants of previous workshops NCGAS via our mailing list. Information was posted via the NCGAS twitter and Facebook page, Indiana University's IT News and Events page as well as the Evolution Directory List Serve. We also contacted The IU Biology department and the XSEDE Campus Champions list directly.

All registration was handled by the IT Training platform. Unlike our workflow-centric workshops, this basic-skills workshop is first come first serve.

Products:

Presentations/websites using modules

2017 Center of Excellence for Women and Technology Research Undergraduate Experience Program

2018 Center of Excellence for Women and Technology Research Undergraduate Experience Program

2019 American Society for Microbiology Workshop – Mining the SRA

"NCGAS/ASM-Workshop-2019-Mining-SRA: ASM Microbe 2019 Workshop 'Using High-Performance Computing (HPC) to mine the NCBI Short Read Archive (SRA).'" [Online]. Available: https://github.com/NCGAS/ASM-Workshop-2019-Mining-SRA. [Accessed: 09-Dec-2019].

2019 National Center for Genome Analysis Support Spring Workshop: Metagenomic Analysis

"NCGAS/Metagenomic-analysis-workshop: The National Center for Genome Analysis Support (NCGAS) offers a three-day workshop on High Performance Computing (HPC) usage and metagenomic analysis between October 7th to 9th, 2019." [Online]. Available: https://github.com/NCGAS/Metagenomic-analysis-workshop. [Accessed: 09-Dec-2019].

Online Course

"SC4E: Intro to R for Biologists." [Online]. Available: https://iu.instructure.com/courses/1766718. [Accessed: 09-Dec-2019].

Online Websites

"R for Biologists." [Online]. Available: https://ncgas.org/R for Biologists Workshop.php. [Accessed: 09-Dec-2019].

Textbook

S. Sanders, *Introduction to R for Biologists*, 1st ed. Bloomington Indiana, 2019.

Virtual Machine

S. Sanders, "Ubuntu18_04 RStudio for NCGAS - Atmosphere Image." [Online]. Available: https://use.jetstream-cloud.org/application/images/871. [Accessed: 09-Dec-2019].

Outcomes:

Before and after our national workshops, we conduct a survey which includes a self-reported confidence level in skills taught in the course. The likert scale is based on the following:

1 No previous experience or knowledge
2 Knowledge of its function, but no hands-on experience
3 Ability to run very limited examples, such as small data sets and tutorials
4 Ability to run more realistic examples, such as real data
5 Ability to troubleshoot tasks for myself and others

The first in-person only instances were used for feedback on material. For our first national run of Introduction to R for Biologists, we taught in a hybrid format. The results of 11 skills ranging from "Using RStudio" to "Writing a custom function" are seen in Table 1.

Table 1. Self-assessed improvement of 11 skills before and after Hybrid Introduction to R for Biologists Workshop. Teaching mode did not significantly impact learning outcome (MANOVA, $p>0.05$)

	All	IUB (hybrid)	IUPUI (in-person)	Online Only
Pre-assessment of skill level	1.91	1.89	1.38	2.05
Post-assessment of skill level	2.99	3.04	2.87	3.02
Improvement	1.08	1.15	1.49	0.97

Groups are not different in pre-comfort level in general (MANOVA, $p>0.05$, N=85). Only installing libraries (lower in in-person at IUB) and installing bioconductor libraries (higher in in-person at IUB) were significantly impacted by access type. There was no post course difference overall or in any individual skill (MANOVA, $p>0.05$, N=43). This was encouraging for our interest in scaling the information to an online format for the national audience.

User reception for this hybrid format was generally positive, though there were a couple of IUB participants that expressed desire for the main instructor to be on site.

For the MOOC style course that represents the end goal of our scaling efforts. The self-reported comfort levels for the same 11 skills were the highest ever in this iteration of the course. This may be a result of the large number of individuals taking the course which will become evident in subsequent runs. Additionally, this form of the course had the lowest post-assessment skill level and the lowest improvement. There was a very limited response to the post-survey which may be driving some of these numbers. From comments in the forum and email, it appears that some individuals appeared hesitant to complete the survey as they did not complete the course. In the future, we will look into possibly building the confidence assessments into the course itself, allowing students to report comfort level after each section.

Table 2. Self-assessed improvement of 11 skills before and after MOOC Introduction to R for Biologists Workshop.

	Online Only
Pre-assessment of skill level (N=183)	2.10
Post-assessment of skill level (N=36)	2.77
Improvement	0.66

Despite this limited reported improvement, the course format appeared to be appealing to the participants. Thirteen of 30 responses to the post-survey question about their favorite part and six of the 29 responses to the most useful part of the course (both free response) was a positive response to the format:

- The self-pacing.
- Multiple ways to get information - YouTube videos, textbook, canvas text, Piazza access. All very appreciated and helpful for learning!

- The workshop was self-paced, because of which I could do the modules in a flexible way. Online lecture videos helped a lot.
- I liked the quizzes embedded in the Modules because it was very satisfying to see that I was gaining experience with commands in R.
- I really liked the format it was in. Even though there were "office hours" the instructors were typically pretty quick to get back to questions. I also ended up reading through the other questions often in case a course mate was having the same issue as I was, which I'm sure saved the instructors time as well. Either way the work at your own pace with help as needed made the course very accessible to me.
- My favorite part was the video lectures. I have attempted to learn R before on my own using manuals and online courses but actually "sitting" in a lecture with a real person teaching, gave me a few "aha" moments that I had never had before.
- The content is more relevant to what I want to do with R in my research, compared to other resource like datacamp. Piazza to share the question across whole participants.
- The best part about the course is that the material is accessible beyond the course timeline. I was not able to complete all the sessions within the two week period, but I'm glad I can continue to learn about R as time permits. Thank you for offering the free course!
- I thought pretty much everything about it was great but it was really nice that students from outside universities could gain this experience without having to pay extravagant course fees.
- I liked that we could see the questions and answers of other colleagues; this served to solve doubts, and the presence of tutors to answer questions.
- Flexibility of workshop schedule and that the course materials are free!!!
- The freedom of the workshop - the pacing was nice. It gave me enough pressure to complete the workshop in a specific window, but did not demand too much of my time. Also, the forum on Piazza was really cool and interesting. I did not actually post a question, but I did read through the posts and found them really helpful.
- The videos were very helpful
- Self-paced!
- Again the piazza forum was great for when I needed help. But also having the course material in the form of an interactive workshop really helped me follow along with and stay interested.
- can ask question at piazza. The materials like textbook is always available at NCGIS website.
- Having access to all materials beyond the class session.
- I found the videos with the printed instructions in the modules very helpful. I could pause the video to type in the code in Rstudio and restart when I was ready. It made it easy to follow along.
- Being available on line

By contrast only 3 of 21 responses to the least useful aspects of the course were in regards to the format (by contrast 5 specifically said they found everything useful!). Three responses were both in regard to the quiz format, which we also found to be a bit awkward. The relevant reponses were:

- Although the quizzes were a good way to test one's understanding and knowledge, for those problems with multiple parts, it was difficult to go between question and answer.

Also, it seemed that in the final chapters there were mistakes in the program instructions and although many were discussed or corrected in the Piazzo chat in some cases it was very difficult to try to find information etc. to find the corrections.

- I was kind of confused about the quizzes because it seemed to me that it was focused around learning new material rather than testing knowledge on old material? That may have been the purpose, but I would have preferred more video lectures in place of the quizzes as someone who is completely new to coding languages.
- I found it a little frustrating trying to troubleshoot issues with code not working etc. and having to go back and forth to the Discussion board.

The quizzes will be revisited in the future, as they were the favorite part of two participants, and allow us to monitor progress through Canvas analytics:

- The whole course was very illuminating. The quizzes allowed for reflection and encouraged further exploration of the concepts through R help topics.
- The quizzes. Quizzes required you to read the documentation for functions before writing code. When I have tried to learn R on my own through manuals, never was the documentation emphasized. It's empowering to understand the documentation.

Workshop Results:

Attendance

Table 3: Registration and Attendance of Each Instance of Workshop

Live Dates:	*Registered*	*Attended*	*Waitlisted*
March - April 2018	17	21 (walk-ins)	0
August - September 2018	32	30	44
November 16, 2018 (TCU)	30	29	0
Hybrid Dates:			
February 2019 - IUB	25	22	0
February 2019 – IUPUI	19	18	0
February 2019 – Online	62	50	6
Online Dates:			
November 2019	330	343 (see note below)	4
Total	*515*	*513*	*48*

Note: This first online course was intended to run for 100 applicants, however, the waitlist became very large, very quickly, and the acceptance list was scaled up. As the online course is open access, some people found that they could take the course without having to register to get around the waitlist. This allowed them access to the material, but not the discussion board (access to which was given to the registered emails).

Engagement Online via Piazza

One of the advantages of online learning management platforms such as Expand/Canvas is that we are able to track engagement in more nuanced way than simply attendance. We confirmed online versions

of the course did not impact learning outcomes, but we also wanted to confirm that the online course version of the course still maintained interaction with the instructors and other students.

We had 223 participants (65% of total) register to use the discussion platform, with 207 (60% of total) of them signing in at least one subsequent day. Participants averaged 3.4 days logged into Piazza, to read an average of 14.15 posts. 160 participants (72% of Piazza, 47% of total) made at least one post (average 1.39 per person), with five people posting more than 10 times in the two weeks.

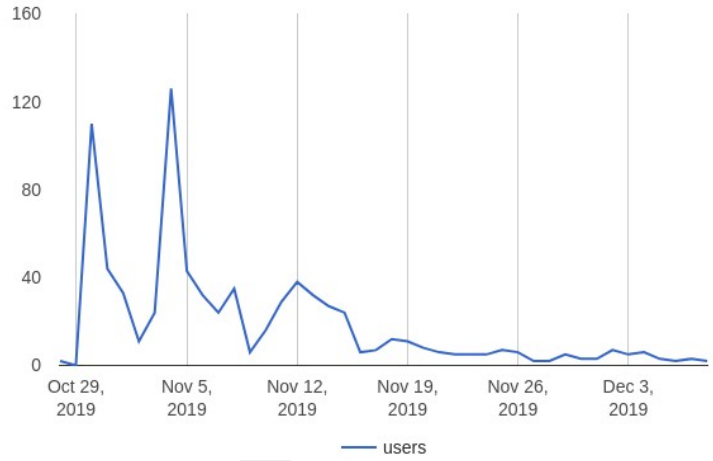

Figure 2. Unique Users Over Time on Piazza Discussion Group, from Piazza Analytics

In total, there were 58 conversation threads (46 started by instructors, 8 initiated by students) over the two weeks, with a total of 429 posts (131 from instructors and 298 were from students). Response to any question over the full two weeks was an average of 12 minutes.

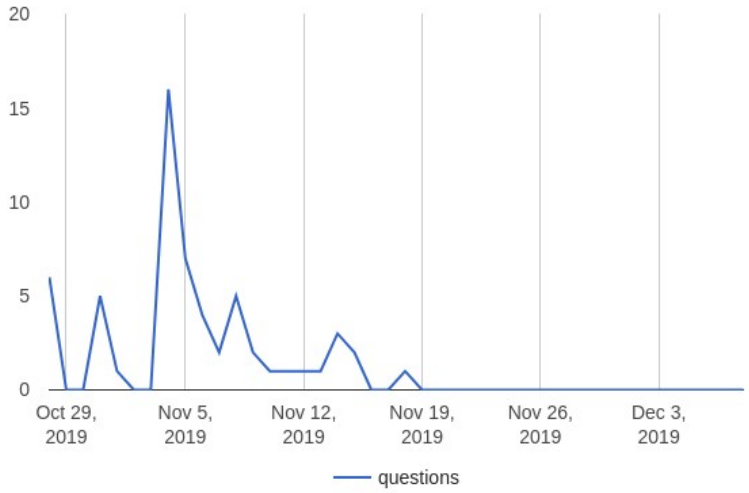

Figure 3. User Questions Over Time on Piazza Discussion Group, from Piazza Analytics

Efficiency of Instructor Time

While ensuring we meet student learning goals and have good engagement are critical aspects of a successful course, scaling courses also requires an efficient decrease in time spent per student enrolled

(while not sacrifcing learning goals or engagement). We were able to steadily reduce the amount of time per student from 1.24 hours to 0.09 hours (almost 14 times more efficient). In addition, the raw instructor time burden increased only 120% while participation increased 1600%.

Table 4: Instructor Time Per Course

Live Dates:	*Prep*	*Course*	*Office Hours*	*Total*	*Hours/Student*
March - April 2018	4hr	6hr*3 inst	4hr *1 inst	26 hr	1.24 hr
August - September 2018	4hr	6hr*3 inst	3hr *1 inst	25 hr	0.83 hr
November 16, 2018 (TCU)	4hr	6hr*2 inst	2hr *1 inst	18 hr	0.6 hr
Hybrid Dates:					
February 2019 - IUB	4hr	9hr*3 inst	6hr *2 inst	43 hr	0.48 hr
February 2019 – IUPUI	-	-	-	-	-
February 2019 – Online	-	-	-	-	-
Online Dates:					
November 2019	4hr	0 (video)	9hrs * 3 inst	31hr	0.09 hr

The most time consuming version of the course to run was the hybrid form of the course, at 43 hours of instructor time. This is a necessary stage, as discussed above, but not the most efficient way to go about teaching large numbers of students. This time did not include driving to separate locations either, which increases the time burden on the staff.

The online course was much easier to manage time wise, as no travel or lectures were required. Instructors were able to answer questions on a more ad hoc schedule, which is more convenient in addition to more time effective. Finally, students did help each other out far more than we saw in in-person courses, likely due to the flexibility of the discussion platform. Finally, as everything is written, adjustments to the course material to include any aspects that were largely unclear is easier than when revisiting the material after verbal conversations with participants.

Future Improvements

- From the post-survey responses, we can see that the quiz structure need to be revisted. We will work with the Canvas team to make these more intuitive.
- We will also make changes to the code that were found to be bugging during this course. We adjusted many as we found them, but all code will be revisited.
- A common request in the “what one thing would you add to the course” and “what other workshops would you like to see us offer in the future” was a request for more specific analyses. Some of the requests were for sepecific analyses that we already have material for (metagenomic analysis, differential expression in R). We will integrate these as further optional modules and continue to add material as we develop it. This will create an online library of R training for our community over time.
- Another common request was to develop similar material for python (planned for late 2020 or 2021) and unix (already in progress). We will make sure to inform past participants when these are available.

<u>Summary of Results and Future Plans:</u>

The National Center For Genome Analysis Support successfully expanded the Introduction to R for Biologists Workshop from a locally offered, 30-person course to a full Massive Open Online Course. The course went through several iterations as an in-person course, then a hybrid course, in order to

solidify materials (including Jetstream Cloud virtual machineds) and record video lectures. This work was done in conjunction with several IU teams.

We were able to show that teaching mode had no effect on learning outcomes, that participation and engagement did not suffer with an online course, and that instructor time was not largely impacted while scaling the course from 30 to 343 students. This course will continue to be offered in person locally (with reduced frequency) and in regular rotation as a MOOC, in order to meet the strong demand for basic coding skills in a domain-guided context.