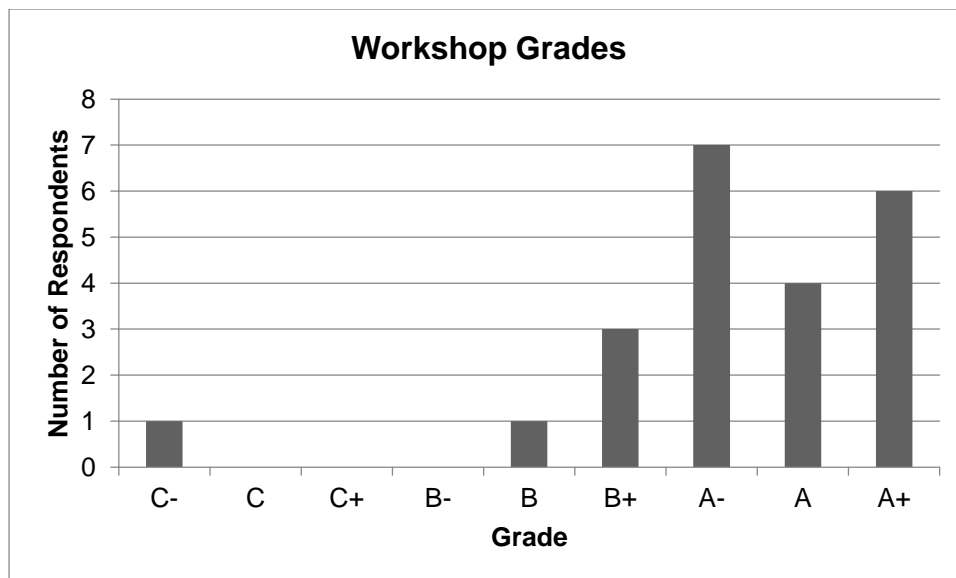


Data Carpentry Evaluation Summary

Summary

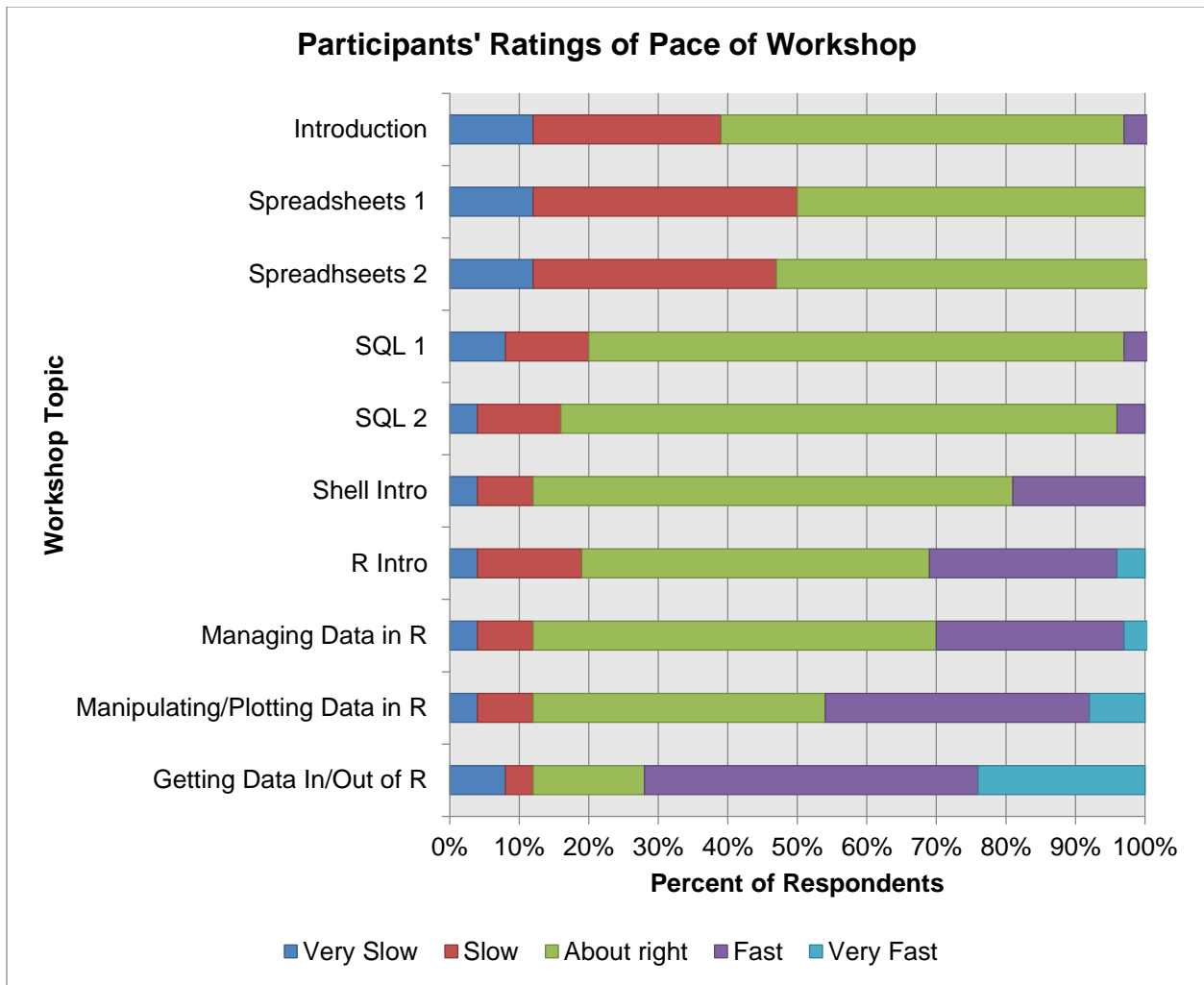
The post-workshop survey was sent to 31 workshop participants; 26 responded yielding an 84% response rate. Nine of the 26 participated remotely, 8 of which did so at the American Museum of Natural History. Approximately one-third of respondents reported their pre-workshop data management and analysis skills as “very low” or “low,” 44% as “neither high nor low,” and 23% as either “high” or “very high.” All but one respondent reported their skills to have increased following the workshop with 33% rating their skills to be “somewhat higher,” 41% “higher,” and 22% “much higher.” All but one respondent reported the workshop to be worth their time, while 84% reported that they would be able to immediately apply what they had learned. Respondents gave the workshop an average grade of A-.



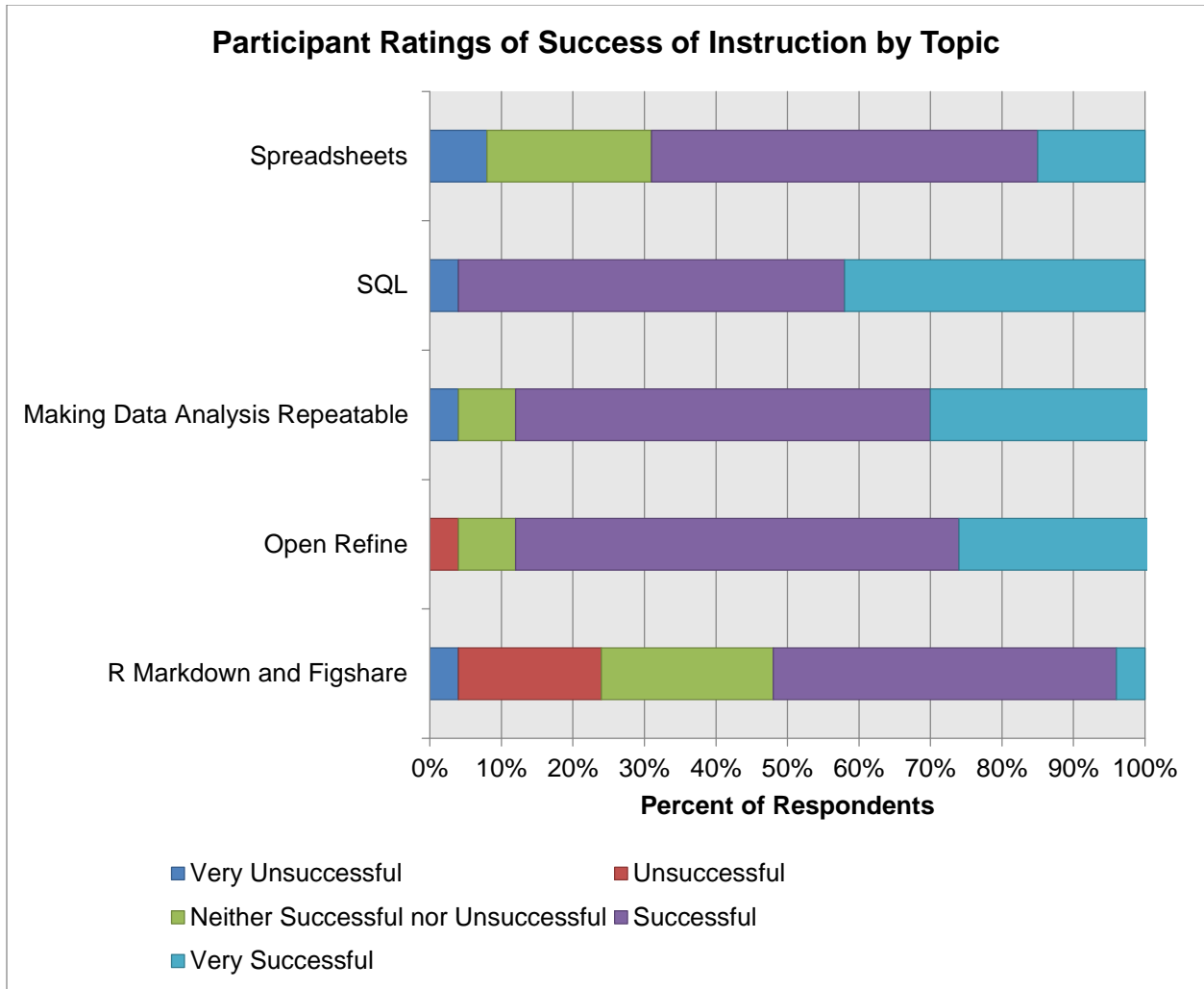


The post-workshop survey included six items also presented on the pre-workshop survey; these items asked respondents about their familiarity with select terms and procedures or how they would solve a data-related problem. Comparisons between pre- and post-responses revealed increased knowledge or skill on each of the six items. For example, following the workshop, 69% of respondents reported they could describe what a database is, when they would use one and how to use it compared to 36% prior to the workshop. The ability to write SQL queries increased from 36% pre-workshop to 85% post-workshop. And the percentage of respondents who could produce a graph based on a dataset using a scripting language almost doubled from 48% prior to the workshop to 92%. Results for each of the six items are shown in Appendix A.

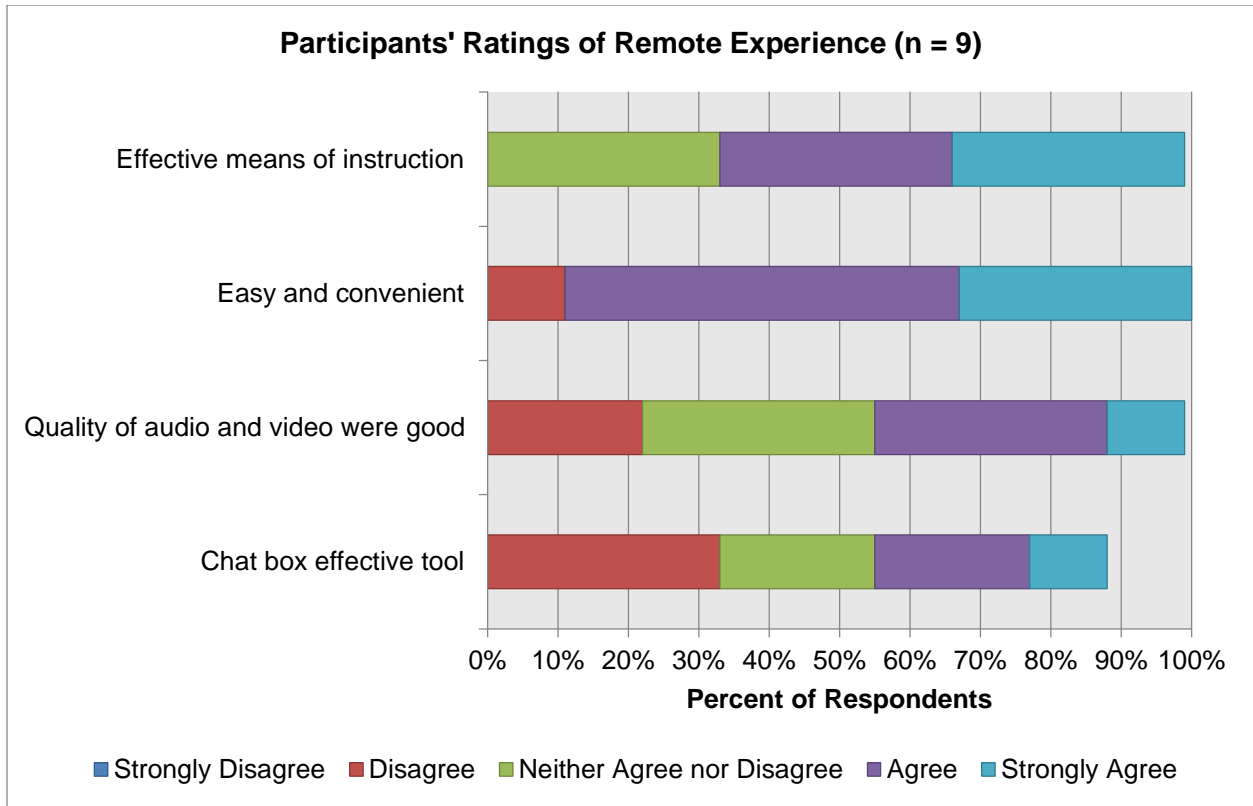
Respondents were also asked to rate the pacing and organization of the workshop. Respondents were most satisfied with the pace of coverage of SQL (more than 75% rated the pace as “about right”), followed by the introduction to shell (69% rated the pace as “about right”), and managing data in R (58% rated the pace as “about right.”). A majority (72%) rated coverage of getting data in and out of R as “fast” or “very fast,” while near majorities (greater than 47%) felt the coverage of spreadsheets as either “very slow” or “slow.”



A majority of respondents rated coverage of all topics as either “successful” or “very successful. The treatment of SQL was perceived as the most successful with 96% rating it as either “successful” or “very successful.” The coverage of R markdown and Figshare was rated as the least successful by far. These ratings are supported by responses to open-ended questions. See Appendix B for verbatim comments about workshop pace and organization.



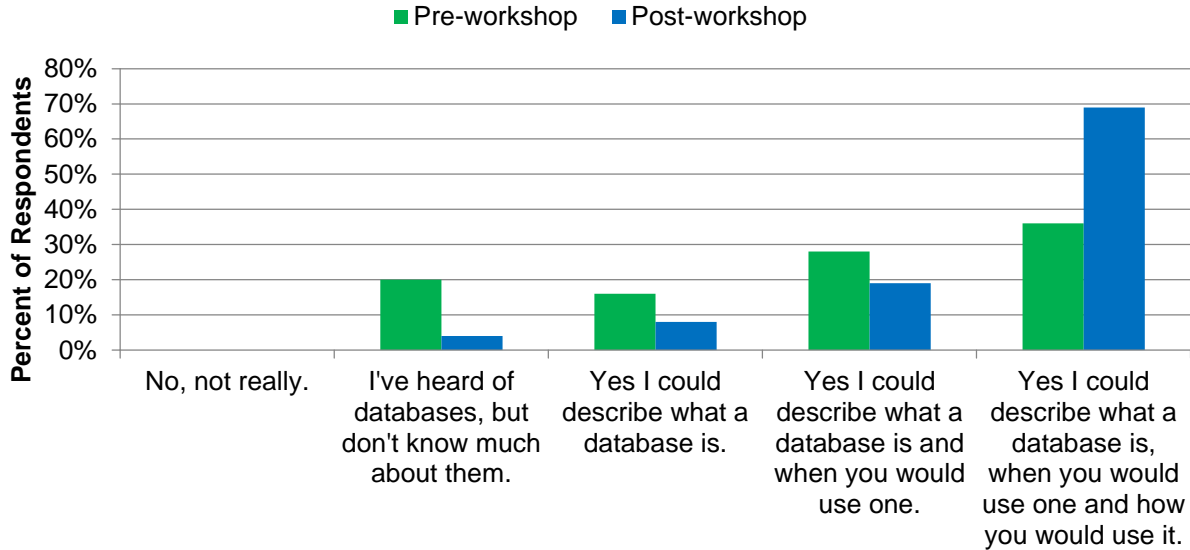
Eight-nine percent of those who participated in the workshop remotely “agreed” or “strongly agreed” that it was easy and convenient to do so. Two-thirds “agreed” or “strongly agreed” that remote participation was an effective means of instruction. There were some technical problems that diminished the quality of the experience for some participants. See Appendix C for comments regarding remote participation.



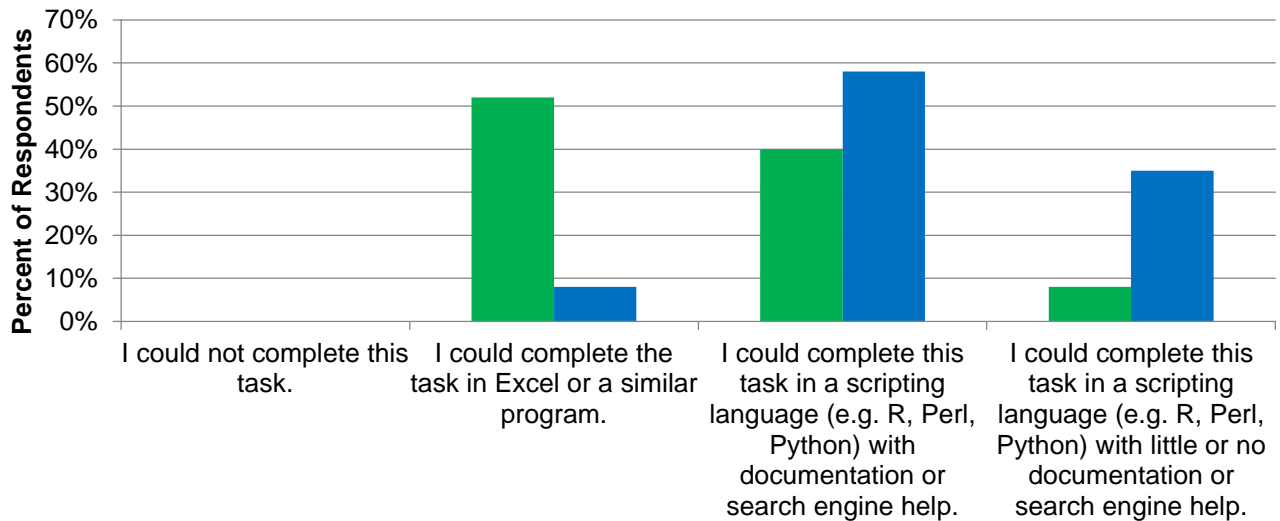
Appendix A. Measures of understanding pre- and post-workshop

These results are based on the responses of 25 individuals who completed both the pre- and post-workshop surveys.

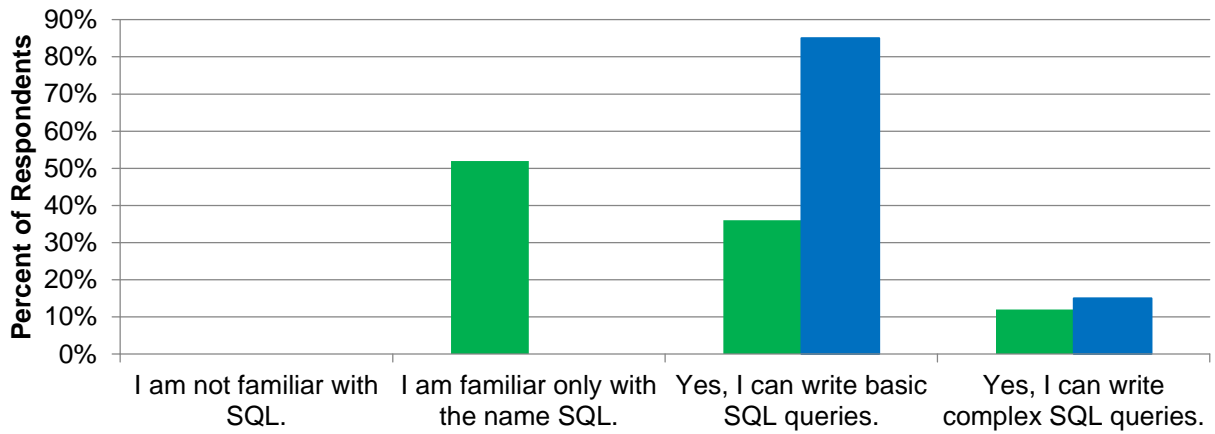
Do you know what a database is and when you might want to use one?



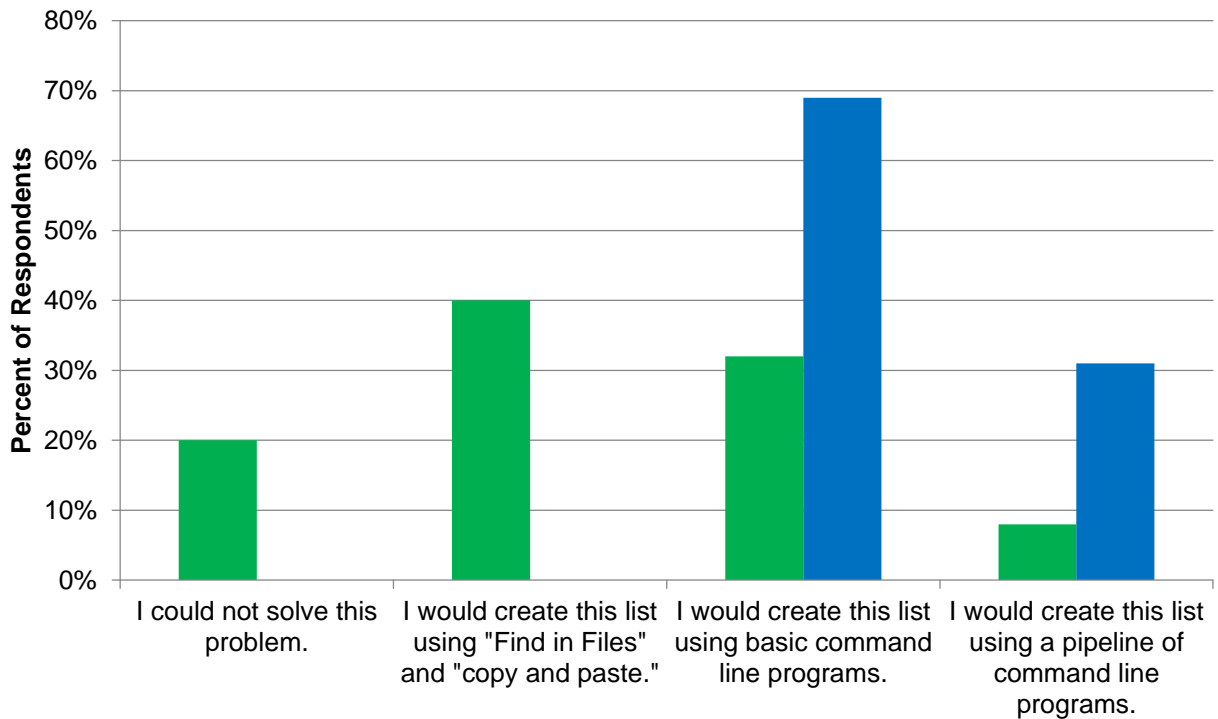
Consider this task: A tab-delimited file has two columns showing the date and the highest temperature on that day. Produce a graph showing the average highest temperature for each month. Which of the following best describes your ability to complete this task?



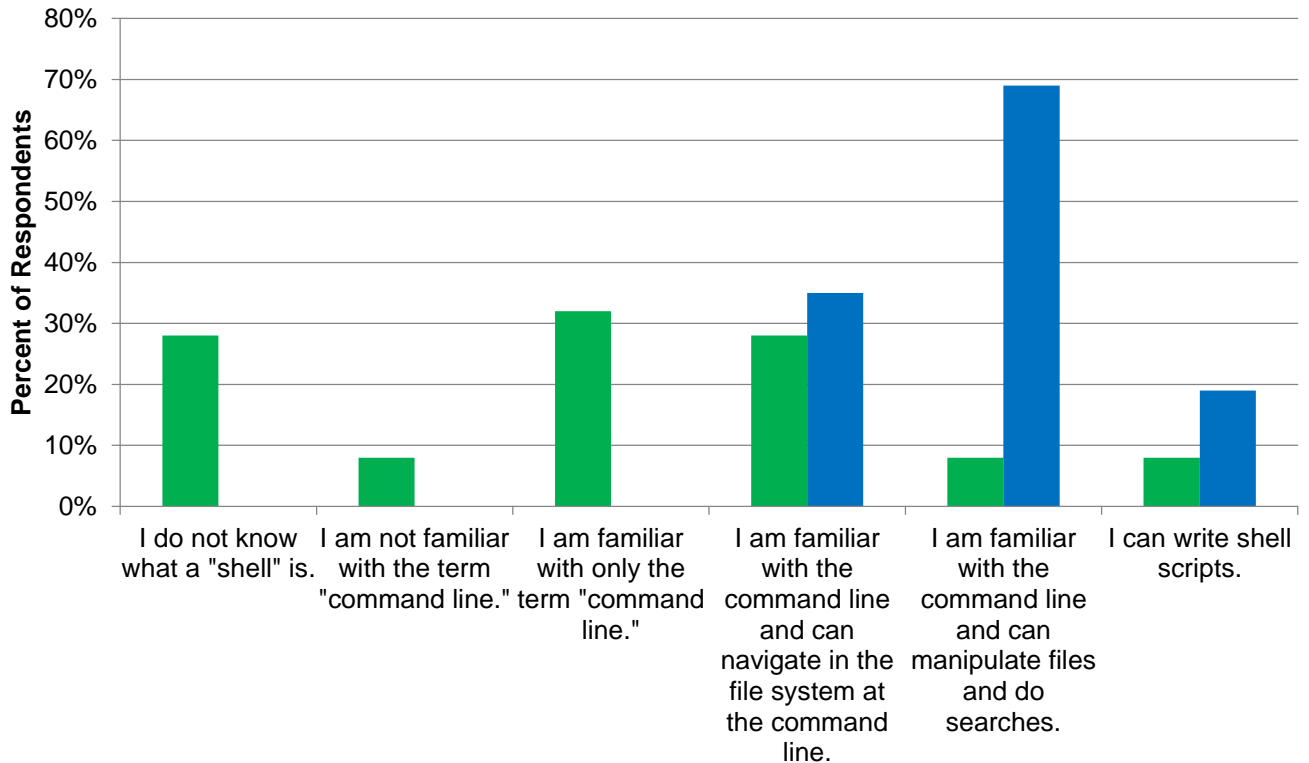
Do you know how to write a SQL query?



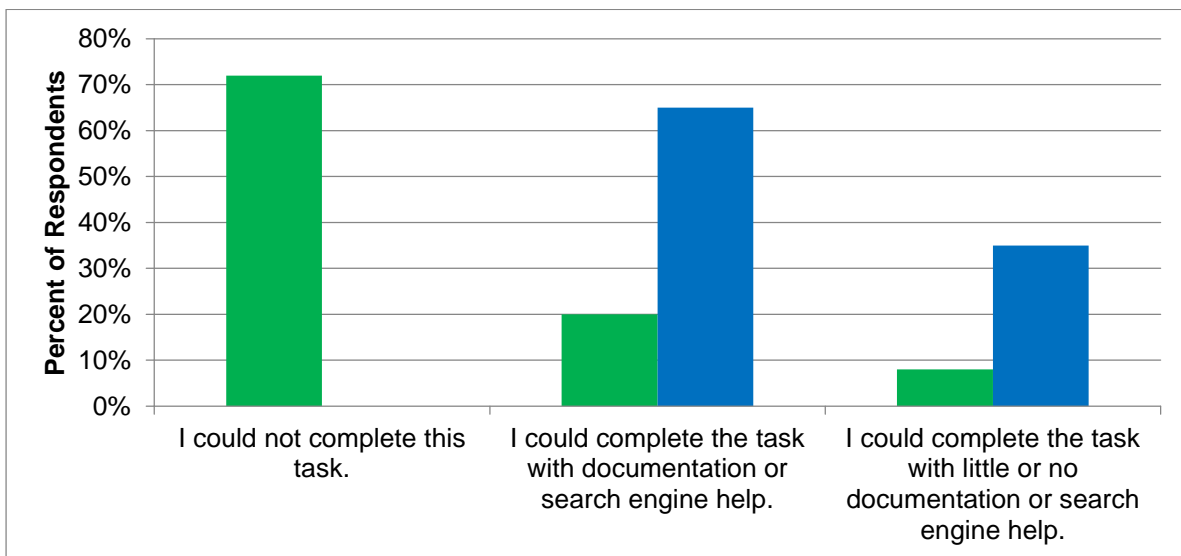
How would you solve this problem? A directory contains 1000 text files. Create a list of all files that contain the word "Drosophila" and save the result to a file called results.txt.



Do you know what a command line is in a shell?



Consider this task: A database has two tables: Scientist and Lab. Scientist's columns are the scientist's user ID, name, and email address; Lab's columns are lab IDs, lab names, and scientist IDs. Write an SQL statement that outputs the number of scientists in each lab.



Appendix B. Ratings and comments on workshop organization and pace

Please share any thoughts you have about the pace or organization of the instruction (n = 21).

Too slow

- Pace was mostly okay. Maybe a tad slow at times.
- I had no background in a few of the concepts/languages that we went over and I still found the intros and exercises a little slow.
- Very slow for what should be typical skill level for graduate student skills. Not enough emphasis on doing. Too many avenues - adobe, etherpad, in person - to coordinate.

Good pace

- Thought we moved at a good pace.
- Awesome organization and pace of the instruction.
- I think the organization was excellent. The pace was good considering the diverse background of the participants. All of the instructors were extremely patient and helpful. I had the impression that everyone was sincerely enthusiastic about the material they were teaching.
- Great overall.

Pace too fast on some topics

- Organization of the instruction (workshop) flowed nicely as each topic led the next one in a unified way. The pace for the introductory sections of the workshop was about right for the most part. For the more complex sections of instruction such as the shell and how to manipulate data in R, the pace was somewhat fast. The fast pace worked if the student had no issues, but if the student had issues because we couldn't understand what was said or technical computer issues with running some of the commands, then that person would have gotten lost since the workshop was moving along fast and since resolving those issues took time and in some instances could not be resolved in the time provided in the workshop. Also the very different level of expertise with computers and programming of the students probably contributed that the pace of more complex sections may have been too fast for those students without much programming experience.
- While I found the workshop very helpful, I thought the R section was pretty rushed. I got lost in the beginning and was never really able to catch up. I think that I would have been happy to sacrifice some of the later R work (or maybe the Shell intro) to have a little more time getting started with R.
- The workshop was great! The only issue, and I'm not sure how to deal with this and include the same amount of instruction, is that when I had problems with R I fell behind while I was trying to figure out what was wrong and then I was behind and running to try to catch up and hope I didn't miss too much.

More R

- This is probably specific to me--but I would have preferred more emphasis on R, and less on SQL and shell. I just feel this way because to me, R has a steeper learning curve than the other two.
- Thanks for a great workshop! Super helpful! R alone is worth 2 full days.
- A brief discussion of how data in R would translate to data in an excel spreadsheet might provide context for the first time R user.
- Less time could have been spent on the introduction so that more time could be spent using R Studio, especially at the end with the workflow, which would have been much more useful. Alternatively, parts of day 1 could have been devoted to introducing R studio so that there was the night to process it for students who were not doing as well with it, and then could come back on day 2 to ask questions before moving to more complex tasks.

The workshop should be longer.

- The workshop was very well organized. Time was maximized. I think one more day would have been useful in order to slow some things down just a bit and also learn some more advanced topics in R.
- I think the program should be extended or the topics should be limited. In other words, only talk about SQL or only talk about R for the two days. I would rather know one of these programs in more detail than both with minimal comprehension. I still don't know what applications I could apply Shell to and I think it may have been better to just remove it entirely and spend additional time on other topics.
- While I understand that it is hard enough to take 2 days off for a workshop, it may be easier in 3 days, and then the pace can be more relaxed.

Other

- Too much time on "why databases" and "dumb things to do in excel"
- GREAT, thanks so much for putting together this.
- very informative, although remote class seems a bit disconnected from onsite discussions.
- I think the workshop was great. It provided a good review on topics I already understood and some basic new things

Please list up to 3 topics for which you would have appreciated more thorough coverage (n = 22).

R (n = 19)

- R (n = 7)
- R workflow (n = 5) *The last section on integrating R into workflow. This section was provided fast because there was not enough time left. And yet it was neat to see the potential of using R scripts to embed in websites and reports.*
- Use of R packages (n = 3)

- R plots (n = 3) *R plotting for phylogenetically or other specific purposes*
- More basic R (for dummies)
- R definitions
- R output
- Functions in R
- Grasping the full potential of R, maybe just some examples. Understand this is beginner stuff though.
- Additional commands/functions, and more examples with data manipulation.
- More R functions for common analyses,
- R scripting
- Databasing in R

SQL (n = 13)

- More SQL (n = 7)
- SQL and the potential communication between R
- SQL and file export
- SQL joint tables and queries
- More practical examples using SQL in collection-based queries
- Have more time with SQL terminology and analyses
- A little more coverage of joins and the different types (at least left, right, inner), and the potential gotcha of a incompletely specified join criteria (cross join). The venn diagram analogy was really helpful to me when learning it on my own:
<http://www.codeproject.com/Articles/33052/Visual-Representation-of-SQL-Joins> (left, right, and inner).

Shell (n = 7)

- More shell (n = 4)
- shell script writing (n = 2)
- Shell: Just a bit of conceptual overview about what the shell actually is. Maybe introduce the distinction between built-ins and executables. A cool example of some kind of data transformation similar to Francios's R markdown example could potentially be inspiring.

Other

- More warnings about where statements instructors made were approximations; most of what they said was 90% right, 90% of the time, but there are plenty of gotchas not mentioned.
- Best practices and basic tips for developing and maintaining datasets and files. Some were mentioned during the first day of the workshop but it would have been good to have a focus 10 or 15 minutes of what those practices are, as the few ones that were mentioned were very valuable.
- A brief introduction to some of the data sharing initiatives or projects. I think this was scheduled for the end of the workshop but we ran out of time.
- Database design

Please list up to 3 topics that should have been de-emphasized in favor of the topics listed above. (n = 20)

Spreadsheet (n = 8)

- Only one session of better spreadsheet setup
- Organizing spreadsheets
- Spreadsheets
- Making the best of spreadsheets--but that's because I have an Information Studies background
- Maybe the spreadsheet orientation?
- Spreadsheets
- The spreadsheet and associated 'tricks' review was helpful, so i hate to say get rid of it... but perhaps that is something that could be reviewed optionally during a lunch break?? or, perhaps a youtube video could be created for it, which could then be reviewed by students prior to the workshop or during a lunch break.

None (n = 6) *Just a little quicker sometimes; Good question -- the agenda was pretty packed, and I really don't know what could potentially be cut to make room for those items.*

Excel (n = 6)

- I cannot list 3. Maybe a bit less excel, but really it is needed.
- Excel pivot tables.
- Excel
- Maybe less excel but I also learned some new stuff with excel during this course so it's hard to say.
- "dumb things to do in excel"
- Basic commands in excel
-

Intro (n = 4)

- "why databases",

Shell (n = 3)

- Basic commands in shell and excel--maybe have a pre-workshop worksheet to get people a little up to speed on these topics?
- Working with shell. (I can't think of a third)
- The shell was interesting but I'm still not sure it is something I would use regularly or ever. I think usefulness of this one may depend in part on the level of expertise the student already has with programming languages, and whether or not they're managing a very large number of files. Perhaps there should be future workshops more geared towards those with programming expertise where things like the shell could be explored in more detail to an audience that really appreciates it. For novices or people without much programming experience, it is hard to convey that importance without first providing some context or a specific example of how the shell helped solved a problem.

R (n = 2)

- Basic coding in R
- Super-complicated R

Given the intended goals, scope and time available for this workshop, are there other topics that should have been discussed? (n = 21)

No (n = 13)

- None (n = 5)
- I was very happy with the knowledge that I took away from this workshop. Thank you.
- No, I don't see how you could have included more than you did.
- I feel a lot more comfortable with programming in general now and that was my primary goal for this workshop, so I definitely consider it a success.
- This is a fantastic introductory workshop. Definitely should not include more topics.
- I don't really see how anything else could be fit in.
- No time for more. This was packed. Good overall choice of coverage.
- I think most of the topics included were quite useful. The issue was the limited time available. Maybe doing a longer workshop or breaking some of the components of the workshop into sub-workshops by themselves may help in the future.
- This workshop was clearly well thought out and reflected a great deal of effort towards preparation. i can't imagine squeezing anything else in. in itself, R could be the topic for the full time. perhaps it's worth offering a 1-day R workshop for people who have NEVER used R, and then that could be a pre-req to the data management workshop. perhaps it could be a self-paced workshop that consists of several you-tube presentations and a 'workbook' (this way students can follow along, pause when they get behind, rewind to a spot of particular interest, etc. I imagine that then it would be easier to hit the ground running during the R portion of the data management workshop.

R (n = 2)

- Just more R.... and maybe providing a bit more context for how R studio functions and the way R 'communicates' with the script writer.
- I would like more R. I have no background in it and I feel like we learned a lot and I definitely want to learn more.

Python (n = 2)

- Python pipeling
- Python scripting

Other

- Other software (e.g.,MS access) and how to import the database to SQL
- Some coverage of the basic commands for moving/copying files should have been covered (e.g. mv cp) to fit with the other elements of what we were doing in shell.
- options for data entry that ensure better quality data.
- best teach yourself resources (books), other courses

Appendix C. Comments on remote participation

Comments about challenges or difficulties participating remotely

- Speaker clarity from host location was fuzzy at times... perhaps have 'stereo surround' mikes/speakers at both locations.
- It was rather difficult to keep up with the many side conversations going on in the other room, especially towards the end with R studio--I'm still not sure why there seemed to be a very long lull towards the end from the instructor when we could have used that time to talk more about workflow since he did not seem to have as good control of the classroom as the others. We were keeping up but unable to hear the conversation about what questions people had.
- Certain voices were easier to hear or more clear to understand than others. I would have liked to save the info in the MoPad but forgot at the time. After I logged in again, I could not save.
- MoPad was best thing. Video/taks were not designed for remote instruction.
- Visual quality low in remote connection.
- Also, instructors should remember to use the computer cursor to direct student attention to the computer screen rather than the projected screen which can't be seen from the remote classroom.

Suggestions on ways to make remote participation more engaging

- Camera MUCH closer to speaker!!!
- It would have been helpful if we were given the data used in the workshop to work with and familiarize ourselves with before the workshop started.
- Review with the instructors the need to repeat the questions and perhaps have a second EtherPad file with the questions being typed into it as they are asked on the screen in AdobeConnect, not separately in Firefox.
- More interaction via chat to not interrupt flow of presenter

Appendix D. Additional Comments

Did the workshop meet your expectations?

Exceeded expectations

- Exceeded!
- Above expectations!
- Exceeded my expectations.. THANKS!!!
- It met all of my expectations and beyond. I didn't expect to follow along with as much ease as I did.

Met expectations

- The workshop met my expectations fully.
- The workshop met my expectations. It was a great workshop. The limiting factor was time both in terms of material covered as there were many important things and also being able to do the hands-on exercises.
- It was in line with my expectations.... I hoped to learn a little more R though.
- Very pleased with my expectations being met
- It was excellent. Learned more than I thought I would.
- Very well (n = 3)
- I really enjoyed the workshop. One of the best I have attended. Very well organized!
- It was exactly what I expected.
- I expected an introduction to the topics, and that is exactly what I got.
- Very well. I didn't expect to become an expert in a couple of days, but I think I now have better tool to learn more on my own.
- definitely.
- It met my expectations

Other

- Enjoyable and informative
- It was helpful to learn a few new things.
- About 8 on a 10 scale.
- Did not meet expectations.

Suggestions for ways to improve future workshops (n = 13)

Provide more information prior to the workshop

- Even more info ahead of time.
- Summary text explaining some of the terms (factor, level, etc), especially in R would be helpful. It would also be good to send a one page overview of concepts and main points for each section ahead of the workshop.
- Providing handouts on definitions of terms that will be mentioned during the presentations.

Limiting focus on topics and/or skill level

- I thought the workshop was well organized. Maybe having different levels of courses would be cool. Beginners, Intermediate, Advanced. Just a thought.
- Simply, limiting the focus might help. I personally would like to know more about SQL and R than other topics discussed.
- Focusing some future workshops on either novices on one hand or experts in programming and data management on the other hand. This may allow to really focus on the very basic or focus on the more complex without either group feeling lost or bored and each group really getting what they are interested in or what they can handle

Address remote challenges

- Better remote set-up.
- This is hard. The remote participation introduces an added factor of technical difficulties; however, it is also nice that it could be a global effort. It would be maybe possible to have trainee instructors, who then could implement the workshop at their home institutions without having to remotely participate.
- Better logistical organization

Other

- I really feel that an extra day would be beneficial. I would have liked to have had the chance to work with my own dataset in R--I think doing hands-on personalized application after learning a topic reinforces what's learned.
- Summary handout on tips and best practices for developing and maintaining datasets
- Possibly save some problem-solving issues for break time.
- More do, less say.

None

- Great job.
- None that comes to mind
- None
- I think the execution of the workshop was excellent. I really don't have any ideas to improve.

Additional Comments

- This workshop should be made available to all undergrads and grad students!
- Katja was a real hero to do AMNH part!
- I would love to help out next time.
- Awesome, please try to do more and more frequent workshop if possible
- Thanks for putting it together.
- Very informative workshop.
- The workshop staff was amazing. Deb did a great job of putting everything together. Matt and Dan were super patient when helping us problem solve. Kevin did a wonderful job with the technology. All of the instructors were truly phenomenal, and special kudos

S. Ellis, November, 2014

to Francois, who not only did a great job teaching and fielding questions about R, but also seemed to genuinely enjoy the process. I look forward to the opportunity to attend another iDig workshop in the future!

- Thanks. The workshop was useful and very informative. Presenters were clear and concise.