DATA 598 A Wi 20: Special Topics In Data Science: Reproducibility for Data Science

Jump to Today

Instructors

Professor Ben Marwick (read about my values, ethics & expectations (http://faculty.washington.edu/bmarwick/values_and_ethics/)

Contact details & office location (http://faculty.washington.edu/bmarwick/#contact)

Office Hours: make an appointment (https://calendly.com/bmarwick)

TA: Ms Liying Wang (liying15@uw.edu (mailto:liying15@uw.edu))

Office Hours: by appointment

Course description

This course, a requirement for the **<u>UW Master of Science in Data Science</u>**

(https://www.washington.edu/datasciencemasters/), introduces students to the principles and tools for computational reproducibility in data science using R. Topics covered include acquiring, cleaning and manipulating data in a reproducible workflow using the tidyverse. Students will use literate programming tools, and explore best practices for organizing data analyses. Students will learn to write documents using R markdown, compile R markdown documents using knitr and related tools, and publish reproducible documents to various common formats. Students will learn strategies and tools for packaging research compendia, dependency management, and containerising projects to provide computational isolation. To complement the learning experience of this class, I recommend subscribing to the email list of the W Reproducible Research Special Interest Group (http://wwescience.github.io/reproducible/) so you can be notified of relevant activities and visiting speakers.

Learning goals

By the end of this course you should be able to:

- 1. Use R to acquire, clean and manipulate data and organise workflows in a reproducible and well-documented way
- 2. Write R Markdown documents that include narrative text, code, and the usual elements of professional writing
- 3. Use Git and GitHub to collaborate on writing reproducible data science
- 4. Make a compendium R package to document and manage dependencies in an analysis
- 5. Use Docker and related tools to provide computational isolation for a data science project, and make it

accessible to others

This is a 2 credit class. We will meet once a week for 2 hours for a combination of lecture, discussion, and hands-on laboratory work. According to the <u>UW guidelines on credit hours</u>
(https://registrar.washington.edu/curriculum/assigning-credit/), you should plan to do 2-4 hours of homework

per week for this class. A calendar view of the course assessments can be seen here.

Schedule of topics

The last day of instruction is Monday 9 March. Any course work due after that date can be completed and submitted remotely. There is no final exam, so you do not need to be on campus after 9 March.

Week	Topic	Reading	R packages
1	Definitions and debates about, and calls for, reproducibility, R, RStudio, Projects (slides (a))	*Barba, L. A. (2018). Terminologies for reproducible research. arXiv preprint arXiv:1802.03311 (https://arxiv.org/abs/1802.03311). Bryan, J. (2017) Project-oriented workflow. Tidyverse Blog. https://www.tidyverse.org/articles/2017/12/workflow-vs-script/ (https://www.tidyverse.org/articles/2017/12/workflow-vs-script/)	here, renv, remotes, pak, fs, sessioninfo, conflicted
2	I OI LI GOMING LI IC	*Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine (https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-7), 8(1), 7. DOI: 10.1186/1751-0473-8-7 (https://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-7) Blischak, John D., Emily R. Davenport, and Greg Wilson. "A quick introduction to version control with Git and GitHub." PLoS Computational Biology (https://journals.plos.org/ploscompbiol/article? id=10.1371/journal.pcbi.1004668) 12, no. 1 (2016): e1004668.	usethis, git2r, gh, gitty,
3	Introduction to R Markdown for documenting a data science project (public	Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. arXiv preprint <u>arXiv:1402.1894</u> (https://arxiv.org/abs/1402.1894). *Baumer, B., & Udwin, D. (2015). R markdown. <u>Wiley.nutredisciplinary Reviews: Computational Statistics</u>	rmarkdown, knitr, rticles, bookdown

	holiday - no slides)	(https://onlinelibrary.wiley.com/doi/full/10.1002/wics.1348), 7(3), 167-177. DOI: 10.1002/wics.1348 (https://onlinelibrary-wiley-com.offcampus.lib.washington.edu/doi/full/10.1002/wics.1348)	
4	Reproducibly manipulating data with the tidyverse, writing functions (slides (a))	*Wickham, H., Averick, M., Bryan, J. et al. (2019). Welcome to the Tidyverse. <i>Journal of Open Source Software</i> (https://joss.theoj.org/papers/10.21105/joss.01686), 4(43), 1686. DOI: 10.21105/joss.01686 (https://joss.theoj.org/papers/10.21105/joss.01686). Wickham, H., & Grolemund, G. (2016). <i>R for Data Science:</i> Import, Tidy, Transform, Visualize, and Model Data. (https://r4ds.had.co.nz/) O'Reilly Media, Inc. https://r4ds.had.co.nz/_(https://r4ds.had.co.nz/).	tidyverse: dplyr, tidyr, purrr, readr
5	Advanced R Markdown documents: external code, caching, templates, etc. (slides 2)	*Xie, Y., Allaire, J. J., & Grolemund, G. (2018). <i>R markdown: The definitive guide (https://bookdown.org/yihui/rmarkdown/)</i> . CRC Press. Chapter 2: Basics (https://bookdown.org/yihui/rmarkdown/basics.html) Xie, Y., Hill, A. P., & Thomas, A. (2017). <i>Blogdown: Creating websites with R markdown.</i> (https://bookdown.org/yihui/blogdown/) Chapman and Hall/CRC.	knitr, redoc, pkgdown, blogdown
6	Packaging data science projects for reproducibility (slides (a))	*Marwick, Ben, Carl Boettiger, and Lincoln Mullen. Packaging data analytical work reproducibly using R (and friends). (https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2017.1375986) The American Statistician 72, no. 1 (2018): 80-88. DOI: 10.1080/00031305.2017.1375986 (https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2017.1375986? journalCode=utas20) Vuorre, M., & Crump, M. J. C. (2020). Sharing and organizing research products as R packages (https://psyarxiv.com/jks2u/). PsyArXiv preprint PsyArXiv:jks2u. DOI: 10.31234/osf.io/jks2u (https://doi.org/10.31234/osf.io/jks2u) Blischak JD, Carbonetto P, and Stephens M. Creating and sharing reproducible research code the workflowr way (https://doi.org/10.12688/f1000research.20843.1) [version 1; peer review: 3 approved]. F1000Research 2019, 8:1749 DOI:	rrtools, workflowr, drake

		10.12688/f1000research.20843.1 (https://doi.org/10.12688/f1000research.20843.1)	
7	Advanced packages for data science projects: testing, documentation, checking (public holiday - no slides)	Wickham, H. (2015). <i>R packages: Organize, test, document, and share your code.</i> (http://r-pkgs.had.co.nz/) O'Reilly Media, Inc. http://r-pkgs.had.co.nz/) (http://r-pkgs.had.co.nz/)	usethis, testthat, tinytest
8	Continuous integration (slides (a))	*Beaulieu-Jones, B. K., & Greene, C. S. (2017). Reproducibility of computational workflows is automated using continuous analysis. (https://www.nature.com/articles/nbt.3780) Nature Biotechnology, 35(4), 342. DOI: 10.1038/nbt.3780	usethis, actions, ghactions, travis, tic. circleci
9	Containerising data science projects (slides 🖹)	*Boettiger, C., & Eddelbuettel, D. (2017). An introduction to rocker: Docker containers for R. (https://journal.r-project.org/archive/2017/RJ-2017-065/index.html) The R Journal 9:2, pages 527-536. Nüst, D. et al. (2020) The Rockerverse: Packages and Applications for Containerization with R (https://arxiv.org/abs/2001.10641). arXiv preprint arXiv:2001.10641 (https://arxiv.org/abs/2001.10641)	containerit, liftr, holepunch
10	Showcasing and archiving projects (slides (a))	*Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. et al. (2017). Our path to better science in less time using open data science tools. (https://doi.org/10.1038/s41559-017-0160) Nature Ecology & Evolution, 1(6), 0160. DOI: 10.1038/s41559-017-0160 (https://doi.org/10.1038/s41559-017-0160) Eglen, S., Marwick, B., Halchenko, Y. et al. Toward standard practices for sharing computer code and programs in neuroscience. Nat Neurosci 20, 770–773 (2017). https://doi.org/10.1038/nn.4550 (https://doi.org/10.1038/nn.4550)	Code Ocean, WholeTale, NextJournal, Zenodo, Figshare

Course assessment and expectations

The course grade consists of the following components and percentages:

• Reading annotations: 30%

• Replication project check-ins: 30%

Replication project: 30%Class participation: 10%

There are some opportunities for extra credit:

- Extra credit via online contributions on CampusWire: 3%
- Extra credit via submission to ReScience C 3%

Your lowest two scores from the *Reading annotations* assignment sets will be dropped from the final grade calculation. This gives you some flexibility to miss some assignments, for example if you have job interviews or busy work periods, without affecting your final grade.

Late submissions: Each student receives three free "late days", each of which allows you to submit an assignment up to 24 hours late without penalty. You will need to notify us by stating on your submitted work that you are using a late day. Once you have used up all late days, assignments will have 10% deducted from the grade per day (including weekends). Assignments will not be accepted more than seven days after the due date (which means you'll get a zero score for that assignment) without prior arrangement. Let me know as soon as you anticipate missing a due date by more than seven days. I review late requests and circumstances on a case by case basis and make decisions accordingly. I generally accommodate unexpected family and medical circumstances, and scheduled religious activities. If you anticipate any disruptions, please let us know so we can help you to plan for success.

You will need a <u>free GitHub account.</u> <u>(https://github.com/join?plan=free&source=pricing-card-free)</u> You may also find other useful things on this <u>UW Libraries Resource Page on Reproducibility</u> (http://guides.lib.uw.edu/research/reproducibility%20).

Communication

We will use <u>CampusWire (https://campuswire.com/c/G54B6BAAE/feed)</u> for announcements and communication about the course. Sign up for an account there with your netid@uw.edu email address. That should give you access—if not, please email me for an access code. The TA and I will post important updates and announcements there. Check often, and be sure to enable your notifications, at least for for the #announcements channel. We recommend you install the Campuswire app on your phone, and ensure that you can receive notifications.

Our Campuswire discussion forum is a shared space for you to ask any class-related questions. Nothing

you post will affect your grade in any way (for better or for worse), so please freely engage by posting questions and by helping your classmates with answers and discussions. Take the time to familiarize yourself with the platform. This is a short introduction to the system: https://www.youtube.com/watch?v=Rz268j1SEq0)

Do post questions about lectures, assignments, etc. You can post a reply or a comment to other students' posts. You can choose to post your question anonymously, although the me and the TA will see your identity. Remember that nothing you post will ever affect your grade! Appropriate, professional images and gifs are welcome. Use the #random channel to share relevant news articles that you come across. Collaborate with your classmates to organise your group project. Use the proper category when posting questions. Use all the tools there that you find helpful. You are welcome to DM me and the TA on Campuswire with questions, comments and feedback about the class. We love to hear from you and value your comments.

Course policies

Grading

The following grading scale will be used:

Percent = Grade

$$90 = 3.5 83 = 2.8 76 = 2.1 69 = 1.4 60-62 = 0.7$$

$$89 = 3.4 \ 82 = 2.7 \ 75 = 2.0 \ 68 = 1.3 < 60 = 0.0$$

Academic misconduct

The university's policy on plagiarism and academic misconduct is a part of the Student Conduct Code, which cites the definition of academic misconduct in the <u>WAC 478-121</u>

(http://www.washington.edu/admin/rules/policies/WAC/478-121TOC.html). (WAC is an abbreviation for the Washington Administrative Code, the set of state regulations for the university. The entire chapter of the WAC on the student conduct code is http://apps.leg.wa.gov/wac/).) According to this section of the WAC, academic misconduct includes:

"Cheating"—such as "unauthorized assistance in taking quizzes", "Falsification" "which is the intentional use or submission of falsified data, records, or other information including, but not limited to, records of

internship or practicum experiences or attendance at any required event(s), or scholarly research"; and "Plagiarism" which includes "[t]he use, by paraphrase or direct quotation, of the published or unpublished work of another person without full and clear acknowledgment."

The UW Libraries have a useful guide for students at http://www.lib.washington.edu/teaching/plagiarism)

Accommodation

Your experience in this class is important to me. If you have already established accommodations with Disability Resources for Students (DRS), please communicate your approved accommodations to me at your earliest convenience so we can discuss your needs in this course. The website for the DRO (https://depts.washington.edu/uwdrs/) provides other resources for students and faculty for making accommodations.

Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW's policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy (https://registrar.washington.edu/staffandfaculty/religious-accommodations-policy/). Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request form (https://registrar.washington.edu/students/religious-accommodations-request/).

Inclusivity

Among the core values of the university are inclusivity and diversity, regardless of race, gender, income, ability, beliefs, and other ways that people distinguish themselves and others. If any assignments and activities are not accessible to you, please contact me so we can make arrangements to include you by making an alternative assignment available.

Learning often involves the exchange of ideas. To include everyone in the learning process, we expect you will demonstrate respect, politeness, reasonableness, and willingness to listen to others at all times – even when passions run high. Behaviors must support learning, understanding, and scholarship.

Preventing violence is a shared responsibility in which everyone at the UW plays apart. If you experience harassment during your studies, please report it to the SafeCampus website (anonymous reports are possible, washington.edu/safecampus/). SafeCampus provides information on counseling and safety resources, University policies, and violence reporting requirements help us maintain a safe personal, work and learning environment.

Absences

In the cases of absences that result in a student missing a course requirement (class activity, assignment submission, exam, e.g.) and of extended absences, accommodations are left to the discretion of the

instructor. Accommodations might include makeup exams, alternate assignments, or alternate weighting of missed work, so long as the grades for other students in the class are not affected by the accommodation.

Technology protocol

A laptop is required in class.

Participation rubric

	Exemplary (90%- 100%)	Proficient (80%-90%)	Developing (70%-80%)	Unacceptable (>70%)
Frequency of participation in class	Student initiates contributions more than once in each class.	Student initiates contribution once in each class.	Student initiates contribution at least in half of the class	Student does not initiate contribution & needs instructor to solicit input.
Quality of comments	Comments always respectful, insightful & constructive; uses appropriate terminology. Comments balanced between general impressions, opinions & specific, thoughtful criticisms or contributions.	Comments mostly respectful, insightful & constructive; mostly uses appropriate terminology. Occasionally comments are too general or not relevant to the discussion.	Comments are sometimes respectful, constructive, with occasional signs of insight. Student does not use appropriate terminology; comments not always relevant to the discussion.	Comments are disrespectful or uninformative, lacking in appropriate terminology. Heavy reliance on opinion & personal taste, e.g., "I love it", "I hate it", "It's bad" etc.
Listening Skills	Student listens attentively when others present materials, perspectives, as indicated by comments that build on others' remarks, i.e., student hears what others say & contributes to the	present ideas, materials, as indicated by comments	Student is often inattentive and needs reminder of focus of class. Occasionally makes disruptive comments while others are speaking	Does not listen to others; regularly talks while others speak or does not pay attention while others speak; detracts from discussion; sleeps, etc.

of focus of comment.

dialogue.

<u>GitHub Organisation for the class (https://github.com/UW-MSDS-DATA-598-Reproducibility-WI20)</u>

<u>Notes on course planning (https://docs.google.com/document/d/1gHTnkOgsf_LGa-</u>

RXbAXzufw7UpPCSnBhtJFSN3NeBJ4/edit)

<u>Instructor/TA meeting notes</u> <u>(https://docs.google.com/document/d/1YZ7x-9ErV-xjWopb3IYBw94hl0axffxh9-SBBD8tzZ8/edit)</u>

Course Summary:

Date	Details	
Mar. Ian 40, 0000	Package notes Week 1 (https://canvas.uw.edu/courses/1354201/assignments/5207772)	due by 23:59
Mon Jan 13, 2020	Reading annotations Week 1 (https://canvas.uw.edu/courses/1354201/assignments/5201507)	due by 23:59
	Participation in Git and GitHub class activity (https://canvas.uw.edu/courses/1354201/assignments/5224865)	due by 23:59
Mon Jan 20, 2020	Project check-in 1: Group formation, identification of target paper and it's key scientific claims, and the key figures to focus on for replication (https://canvas.uw.edu/courses/1354201/assignments/5218683)	due by 23:59
	Reading annotations Week 2 (https://canvas.uw.edu/courses/1354201/assignments/5203541)	due by 23:59
Mon Jan 27, 2020	Project check-in 2: Identification of data, tools and infrastructure necessary (https://canvas.uw.edu/courses/1354201/assignments/5218685)	due by 23:59
,	Reading annotations Week 3 (https://canvas.uw.edu/courses/1354201/assignments/5203546)	due by 23:59
M 5 1 0 0000	Participation in Intro R to Markdown class activity (https://canvas.uw.edu/courses/1354201/assignments/5224873)	due by 23:59
Mon Feb 3, 2020	Reading annotations Week 4 (https://canvas.uw.edu/courses/1354201/assignments/5209722)	due by 23:59
Mon Feb 10, 2020	Participation in Advanced R Markdown class activity (https://canvas.uw.edu/courses/1354201/assignments/5244870)	due by 23:59

	Reading annotations Week 5 (https://canvas.uw.edu/courses/1354201/assignments/5209724)	due by 23:59
Mon Feb 17, 2020	Participation in Basic Research Compendium Package class activity (https://canvas.uw.edu/courses/1354201/assignments/5253071)	due by 23:59
	Reading annotations Week 6 (https://canvas.uw.edu/courses/1354201/assignments/5209730)	due by 23:59
Mon Feb 24, 2020	Project check-in 3: Replication of the key figures (https://canvas.uw.edu/courses/1354201/assignments/5218687)	due by 23:59
WOT FED 24, 2020	Reading annotations Week 7 (https://canvas.uw.edu/courses/1354201/assignments/5209732)	due by 23:59
Mon Mar 2, 2020	Participation in Advanced R Packages class activity (https://canvas.uw.edu/courses/1354201/assignments/5270867)	due by 23:59
WOT Wat 2, 2020	Reading annotations Week 8 (https://canvas.uw.edu/courses/1354201/assignments/5209736)	due by 23:59
	Participation in Containerisation class activity (https://canvas.uw.edu/courses/1354201/assignments/5279011)	due by 23:59
	Participation in Replication Project presentation reviews class activity (https://canvas.uw.edu/courses/1354201/assignments/5281651)	due by 23:59
Mon Mar 9, 2020	Project check-in 4: Peer review (https://canvas.uw.edu/courses/1354201/assignments/5218689)	due by 23:59
	Project check-in 5: In class presentation (https://canvas.uw.edu/courses/1354201/assignments/5218691)	due by 23:59
	Reading annotations Week 9 (https://canvas.uw.edu/courses/1354201/assignments/5209740)	due by 23:59
Mon Mar 16, 2020	Reading annotations Week 10 (https://canvas.uw.edu/courses/1354201/assignments/5209743)	due by 23:59
WOTT WAT 10, 2020	Replication project final submission (https://canvas.uw.edu/courses/1354201/assignments/5218657)	due by 23:59
	Credit-workload adjustment (https://canvas.uw.edu/courses/1354201/assignments/5290332)	
	In-class participation (https://canvas.uw.edu/courses/1354201/assignments/5218708)	
	Online participation	

(https://canvas.uw.edu/courses/1354201/assignments/5218667)



Submission to ReScience C (https://canvas.uw.edu/courses/1354201/assignments/5218671)



(http://creativecommons.org/licenses/by/4.0) This course content is offered under a <u>CC Attribution</u> (http://creativecommons.org/licenses/by/4.0) license. Content

in this course can be considered under this license unless otherwise noted.