Teaching reproducibility and responsible workflows: an educator and editor's perspective

Nicholas J. Horton, Amherst College

August 8, 2023, JSM, nhorton@amherst.edu

Image source: heylagostechie

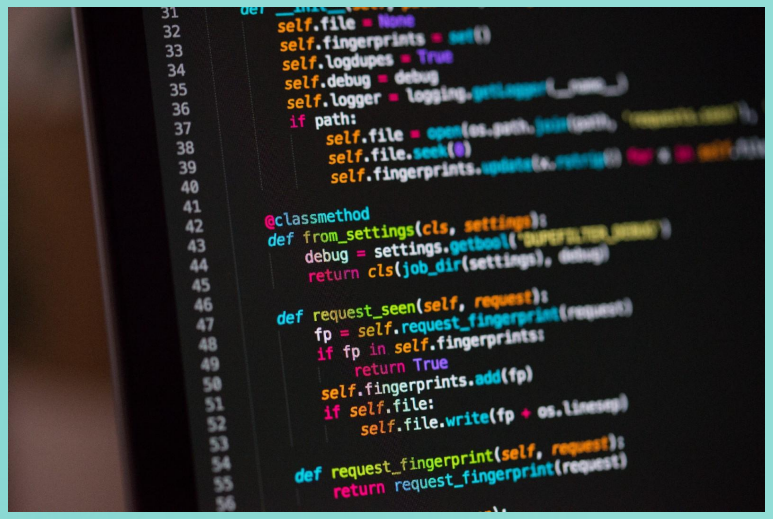


Image source: Wikicommons

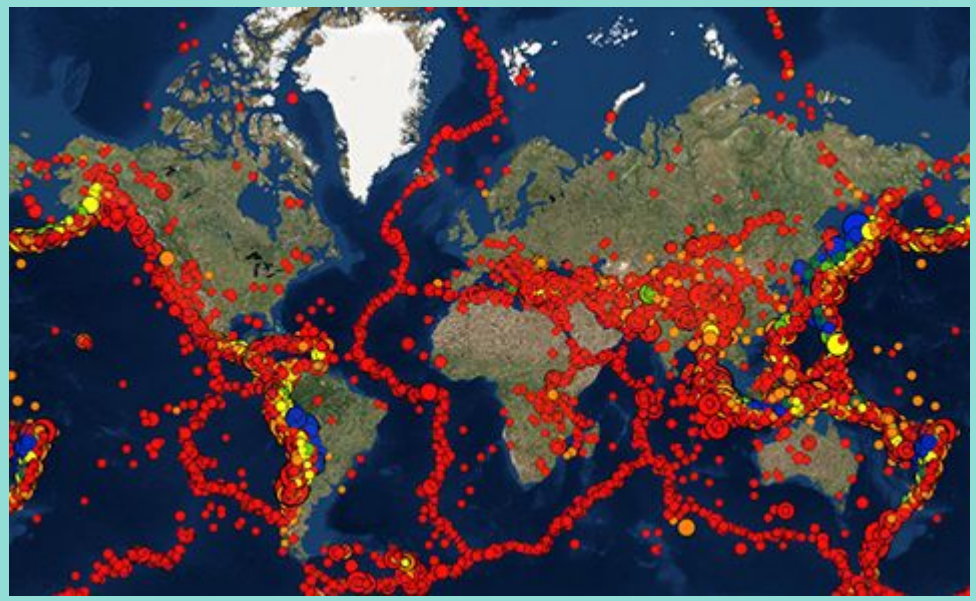


Image source: Concord Consortium

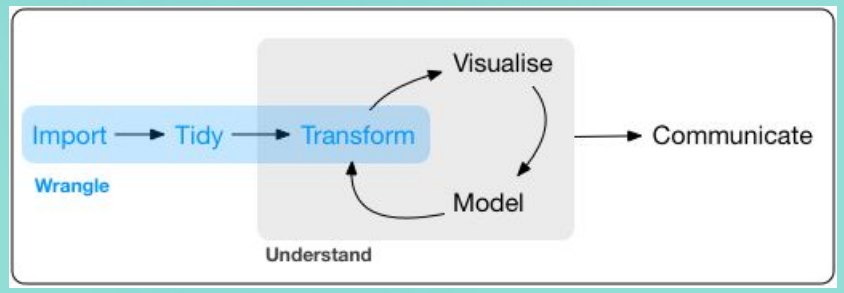


Image source: Hadley Wickham and Garrett Grolemund

thanks to NSF #1923388

Acknowledgements

- Many key ideas derive from my collaborators, including Valerie Barr, Ben Baumer, Matt Beckman, Mine Çetinkaya-Rundel, Jie Chao, Bill Finzer, Jo Hardin, Chelsey Legacy, Randall Pruim, Maria Tackett, and Andy Zieffler

The growing importance of reproducibility and responsible workflow in the data science and statistics curriculum

https://www.tandfonline.com/toc/ujse21/30/3?nav=tocLis

special issue (November 2022) of the *Journal of Statistics and Data Science Education,*

Aneta Piekut
Univ. of Sheffield

Colin Rundel
Duke University

Micaela Parker
ADSA

Nicholas Horton
Amherst College

Rohan Alexander
Univ. of Toronto

JSDSE special issue (November 2022)

- "The growing importance of reproducibility and responsible workflow in the data science and statistics curriculum" (Horton et al, https://doi.org/10.1080/26939169.2022.2141001)
- "An invitation to teaching reproducible research: lessons from a symposium" (Ball et al https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2099489)
- "Interdisciplinary approaches and strategies from research reproducibility 2020: educating for reproducibility" (Rethlefsen et al, , https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2104767)
- "Data science ethos lifecycle: interplay of ethical thinking and data science practice" (Boenig-Liptsin et al, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2089411)
- "Opinionated practices for teaching reproducibility: motivation, guided Instruction and practice" (Ostblom and Timbers, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2074922)
- "Tools and Recommendations for Reproducible Teaching" (Dogucu and Çetinkaya-Rundel, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2138645)
- "Third Time's a Charm: A Tripartite Approach for Teaching Project Organization to Students" (Mehta and Moore, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2118644)

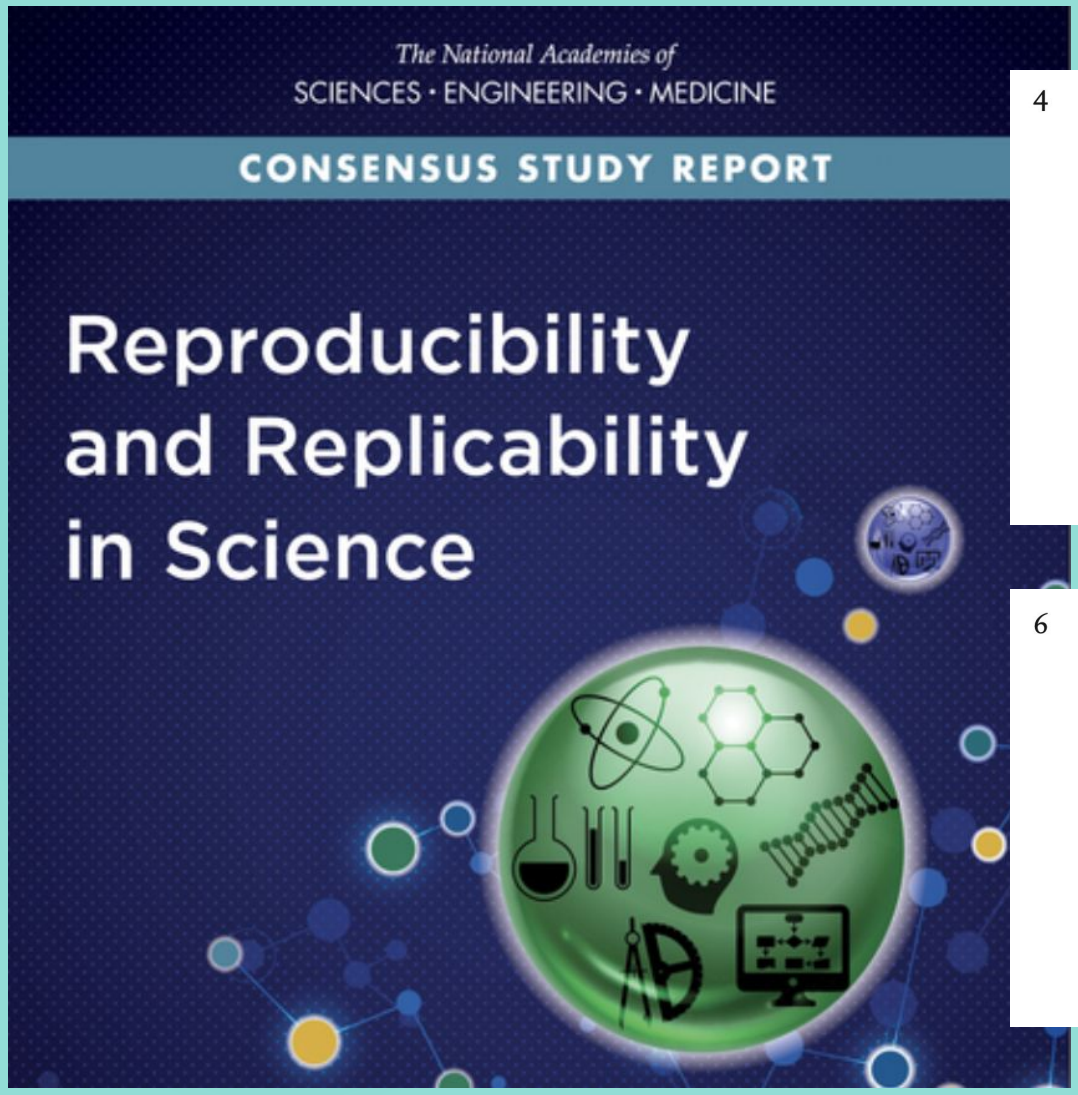
JSDSE special issue (November 2022)

- “LUSTRE: An online data management and student project resource” (Towse et al,, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2118645)
- “Teaching for Large-Scale Reproducibility Verification” (Vilhuber et al, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2074582)
- “Collaborative Writing Workflows in the Data-Driven Classroom: A Conversation Starter” (Sara Stoudt, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2082602)
- “A Journey from Wild to Textbook Data to Reproducibly Refresh the Wages Data from the National Longitudinal Survey of Youth Database” (Amaliah et al,, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2094300)
- “Approachable case studies support learning and reproducibility in data science: An example from evolutionary biology” (Sanchez Reyes and McTavish https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2099487)

Plan

- The importance of reproducibility when teaching data acumen
- Initiatives to teach reproducibility and responsible research (making change happen)
- An editor's perspective on next steps to foster reproducible practice and research

NASEM (2019)



4 REPRODUCIBILITY
Widespread Use of Computational Methods, 55
Nonpublic Data and Code, 57
Resources and Costs of Reproducibility, 57
Assessing Reproducibility, 59
The Extent of Non-Reproducibility, 62
Sources of Non-Reproducibility, 67
Inadequate Recordkeeping, 67
Nontransparent Reporting, 69
Obsolescence of Digital Artifacts, 69
Flawed Attempts to Reproduce Others' Research, 70
Barriers in the Culture of Research, 70

6 IMPROVING REPRODUCIBILITY AND REPLICABILITY
Strengthening Research Practices: Broad Efforts and Responsibilities, 105
Education and Training, 108
Improving Knowledge and the Use of Statistical Significance Testing, 109
Efforts to Improve Reproducibility, 110
Recordkeeping, 111
Source Code and Data Version Control, 114
Scientific Workflow-Management Systems, 114
Tools for Reproduction of Results, 116
Publication Reproducibility Audits, 118

NASEM (2019)

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey **clear, specific, and complete information** about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis, unless such information is restricted by nonpublic data policies.

That information should include the data, study methods, and computational environment:

NASEM (2019)

1) the input data used in the study either in extension (e.g., a text file or a binary) or in intension (e.g., a script to generate the data), as well as intermediate results and output data for steps that are nondeterministic and cannot be reproduced in principle;

2) a detailed description of the study methods (ideally in executable form) together with its computational steps and associated parameters; and

3) information about the computational environment where the study was originally executed, such as operating system, hardware architecture, and library dependencies.

NASEM (2019)

- RECOMMENDATION 6-6: Many stakeholders have a role to play in improving computational reproducibility, including educational institutions, professional societies, researchers, and funders.
- **Educational institutions should educate and train students and faculty about computational methods and tools to improve the quality of data and code and to produce reproducible research.**
- Professional societies should take responsibility for educating the public and their professional members about the importance and limitations of computational research.

Problem-solving cycle

- What are we hoping that students will learn?
- How can tools for reproducibility help scaffold their learning?

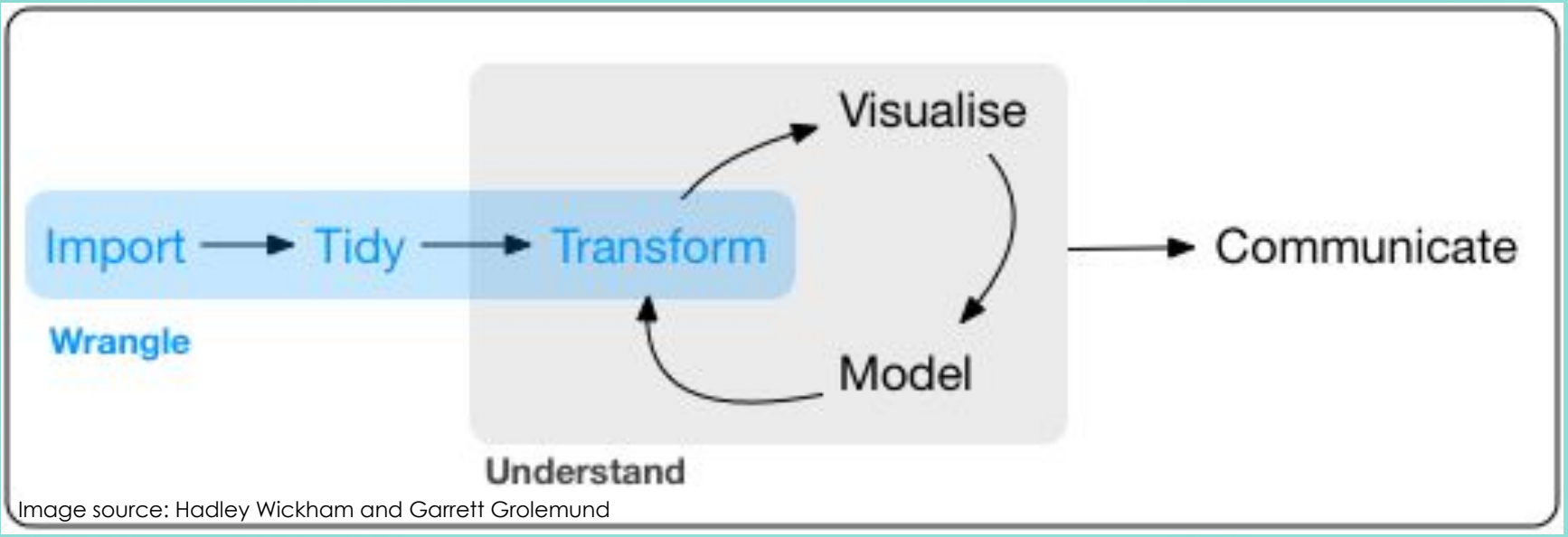


Image source: Hadley Wickham and Garrett Grolemund

DATA SCIENCE FOR UNDERGRADUATES

Opportunities and Options

consensus report published in 2018
https://nas.edu/envisioningds

Study funded by the National Science Foundation

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Key Insights NASEM (2018): Undergraduate Data Science

- There must be **multiple pathways** for undergraduates to study data science
- The undergraduate experience should cater to and **promote diversity** – demographic and intellectual – in the students it serves
- There are some core competencies that all data science students (and, ideally, all undergraduates) should have
  - They should develop **data acumen**
  - Ethical problem-solving is a key component of data acumen

A Central Finding

Finding 2.3 A critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved in developing data acumen include the following:

- Mathematical foundations
- Computational foundations
- Statistical foundations
- Data management and curation
- Data description and visualization
- Data modeling and assessment
- Workflow and reproducibility
- Communication and teamwork
- Domain-specific considerations
- Ethical problem solving.

A Central Finding

Finding 2.3 A critical task in the education of future data scientists is to instill data acumen. This requires exposure to key concepts in data science, real-world data and problems that can reinforce the limitations of tools, and ethical considerations that permeate many applications. Key concepts involved in developing data acumen include the following:

- Mathematical foundations
- Computational foundations
- Statistical foundations
- **Data management and curation**
- **Data description and visualization**
- Data modeling and assessment
- **Workflow and reproducibility**
- **Communication and teamwork**
- Domain-specific considerations
- **Ethical problem solving.**

Bolded areas indicate direct connections with reproducibility as defined broadly

Computational concepts

While it would be ideal for all data scientists to have extensive coursework in computer science, new pathways may be needed to establish appropriate depth in **algorithmic thinking and abstraction** in a streamlined manner. This might include the following:

- Basic abstractions,
- Algorithmic thinking,
- Programming concepts,
- Data structures, and
- **Simulations**.

Idea of "computational essay"

Data modeling concepts

Key **data modeling and assessment** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

Idea of "computational essay"

- Machine learning,
- Multivariate modeling and supervised learning,
- Dimension reduction techniques and unsupervised learning,
- Deep learning,
- Model assessment and sensitivity analysis, and
- Model interpretation (particularly for black box models).

Data management concepts

Key **data management and curation** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- **Data provenance;**
- **Data preparation, especially data cleansing and data transformation;**
- **Data management (of a variety of data types);**
- Record retention policies;
- Data subject privacy;
- **Missing and conflicting data**; and
- Modern databases.

information/library science skills

Workflow and reproducibility concepts

Key **workflow and reproducibility** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- **Workflows and workflow systems,**
- **Reproducible analysis,**
- **Documentation and code standards,**
- **Source code (version) control systems, and**
- **Collaboration.**

archival skills

Communication and teamwork concepts

Key **communication and teamwork** concepts/skills that would be important for all students in their data science programs and critical for their success in the workforce are the following:

- Ability to understand client needs,
- **Clear and comprehensive reporting,**
- **Conflict resolution skills,**
- Well-structured technical writing without jargon, and
- Effective presentation skills.

Ethical concepts

Key aspects of **ethics** needed for all data scientists (and for that matter, all educated citizens) include the following:

- **Ethical precepts for data science and codes of conduct,**
- **Privacy and confidentiality,**
- **Responsible conduct of research,**
- Ability to identify "junk" science, and
- Ability to detect algorithmic bias.

The importance of version control

- We've already heard how systems such as git and GitHub allow individuals and groups to document changes to files over time
- Improved version control can improve collaboration
- Or help communicate with ourselves six months in the future!
- Caveat: expert friendly

Beckman et al (JSDSE, 2021)

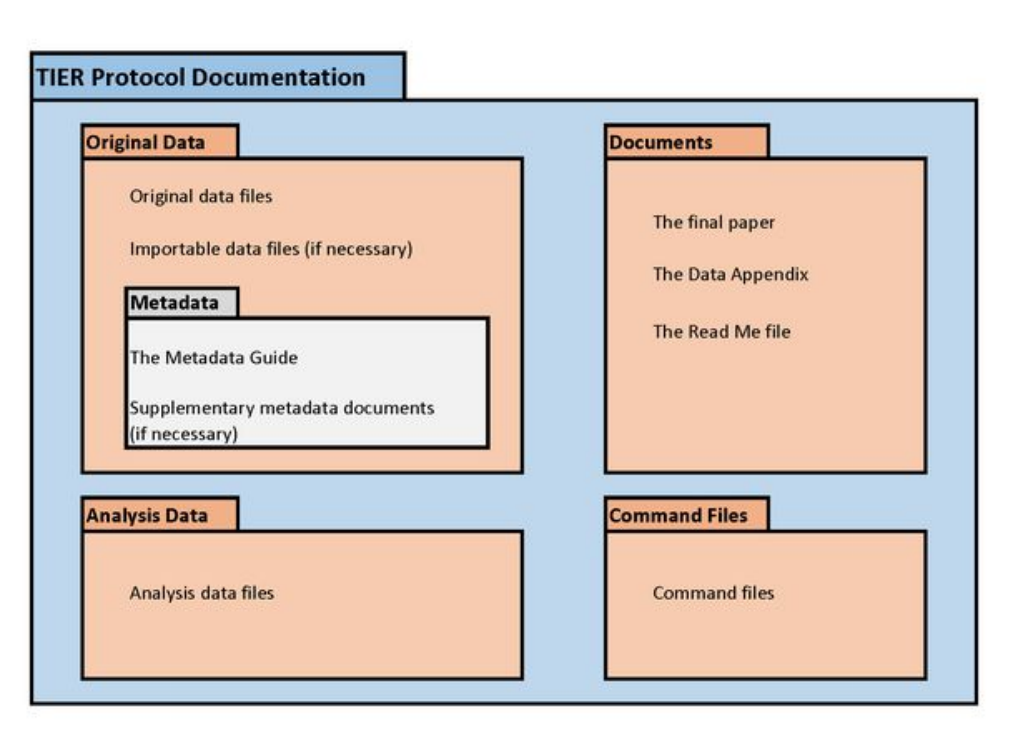
- Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses
- https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1848485
- Student experiences from Amherst College, Brown University, Duke University, and the University of Edinburgh
- "Teaching reproducible analysis in the statistics curriculum helps make students aware of the issue of scientific reproducibility and also equips them with the knowledge and skills to conduct their future data analyses reproducibly, whether as part of an academic research project or in industry."

Beckman et al (JSDSE, 2021)

- Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses
- https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1848485
- "Use of version control helps reinforce the notion that statistical analysis typically requires multiple revisions, as students can review their commits to see all the updates they've made to their work. A desirable side effect is that, because students are periodically "submitting" their assignment as they work on it, there is less pressure of the final deadline where everything must be submitted in its final form."

Foundational work: TIER protocol 3.0

- https://www.projecttier.org
- At first focused on substantial research projects
- Now working to build a developmental progression across the undergraduate curriculum (see the soup to nuts exercises)



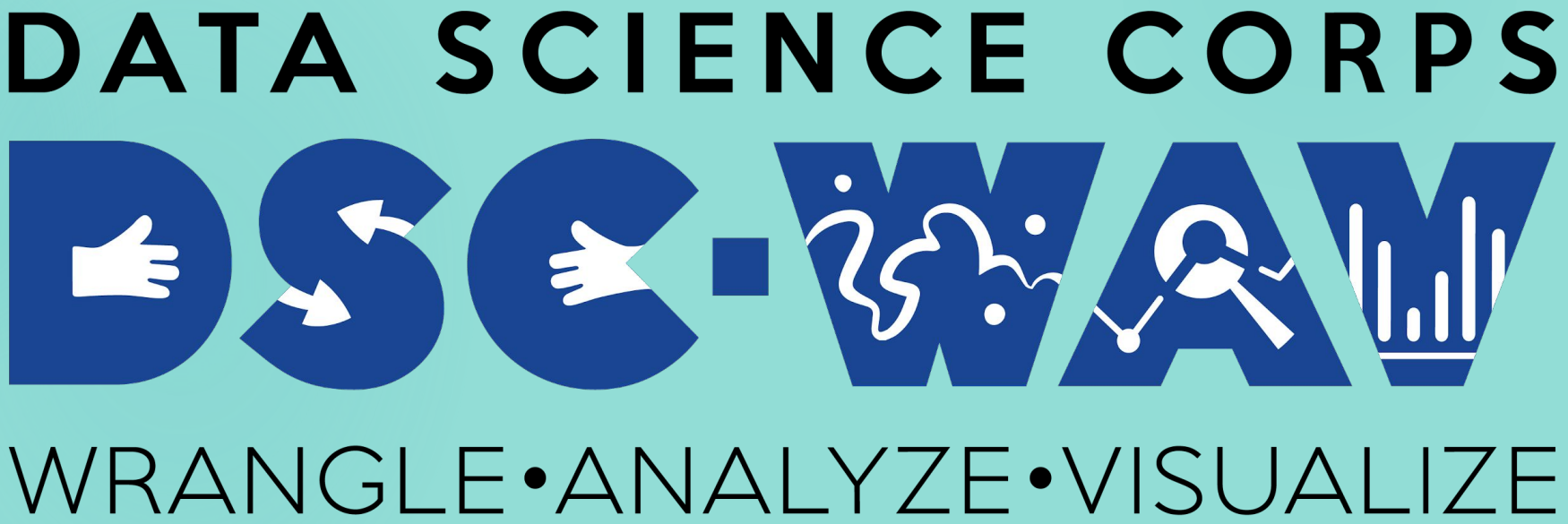
TIER symposium

- creative ten-part virtual event, two part presentation + Q&A
- March 5 - May 21, 2021
- Passover/Easter thinking gap
- “slow food” metaphor well-suited to the pandemic

Instruction in Reproducible Research: Educational Outcomes

DSC-WAV (Wrangle-Analyze-Visualize)

- NSF funded effort from the Harnessing the Data Revolution (HDR) Data Science Corps (DSC) initiative: https://dsc-wav.github.io/www



Agile and scrum for undergraduates

- "Facilitating team-based data science: Lessons learned from the DSC-WAV project", *Foundations of Data Science* (Legacy et al, https://www.aimsciences.org/article/doi/10.3934/fods.2022003

The inspiration for the DSC-WAV program was a question of whether undergraduate students could tackle real-world data science problems utilizing the tools and approaches frequently seen in industry. Based on our experiences, the answer to this question is "yes."

Source: smartbear.com

Source:Esti Alvarez, see also https://teachdatascience.com/pairprogramming

Key questions

What are best practices for teaching these methods?

- Very much a work in progress
- Developmental progression is important (beginning in K-12)
- Need to develop, pilot, improve, then vet course materials and curricular modules
- Need to bridge gap between teaching and practice
- More work is needed!

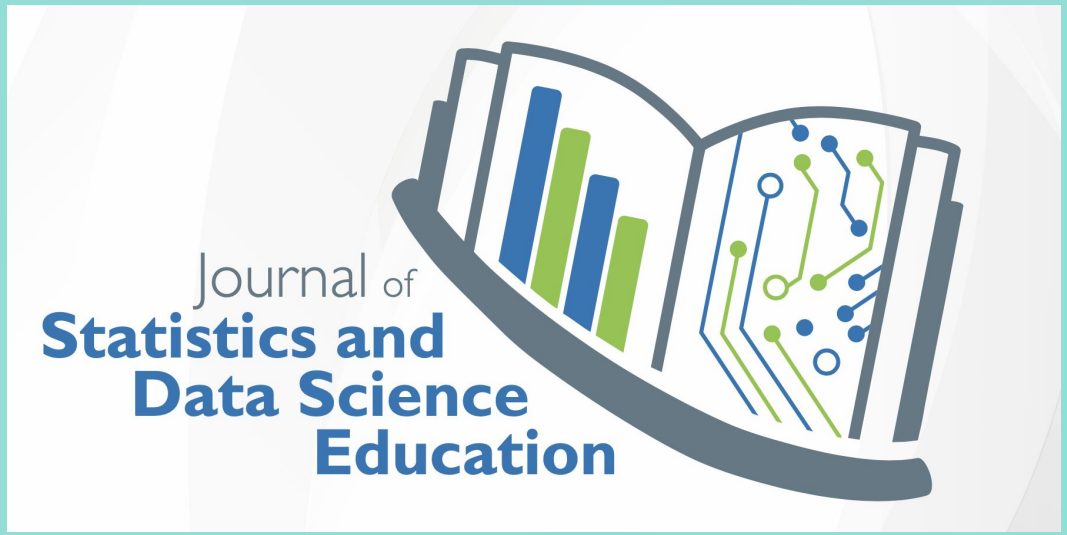
Key questions

Where can these methods be incorporated into the K-12 and college curricula?

- Lab notebooks and science curriculum in K-12 (see CODAP and Concord Consortium)
- Early and often in college: Bussberg (TIER) argued that reproducibility should be in all courses to avoid:
  - developing bad habits
  - reinforcing good habits
  - building skills needed for workforce and graduate studies
- Feeds into innovative approaches like Janz, Sullivan, and McAleer courses (see also JSDSE special issue)

Editor's perspective: making change happen

- The *Journal of Statistics and Data Science Education* (formerly *Journal of Statistics Education*) is published by Taylor & Francis on behalf of the American Statistical Association.
- Open access with no author fees
- Submissions on reproducibility and workflow (and other topics) welcomed, https://www.tandfonline.com/loi/ujse21



JSDSE next steps: Data and code sharing

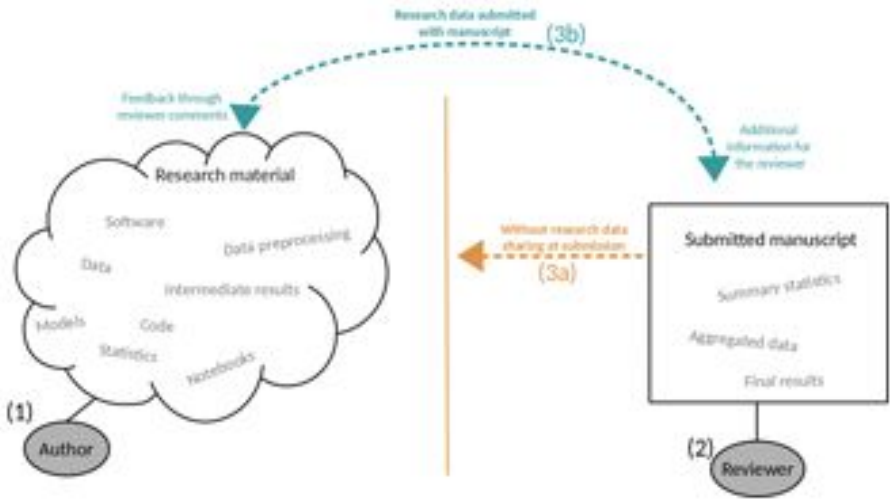
When should data and code be made available?

Significance Magazine (April, 2022)

Rachel Heyard, Leonhard Held

Pages: 4-5 | First Published: 29 March 2022

Sharing data and code as part of a research publication is crucial for ensuring the computational reproducibility of scientific work. But sharing should be done at the article submission stage, not after publication as it is now, say Rachel Heyard and Leonhard Held. Statisticians and data scientists have the skills and tools to make this change and lead by example, though there are obstacles to overcome.



As of September 1, 2022, all submissions to the *Journal of Statistics and Data Science Education* require a "Data Availability Statement" which outlines how code and data underlying a paper have been made available.

See https://www.tandfonline.com/journals/ujse21 for more info

JSDSE next steps: Data and code sharing

- Goal: foster better reproducible science by adopting open science frameworks
- Precedent: JASA established guidelines in 2016
- Decision: Data and code sharing requirements added as of September 1, 2022
- Implementation: Authors deposit their code and data in an online repository (e.g., OSF.io) and include a link in the article metadata
- Authors also include a Data Availability Statement just before the references
- The link to data are code are provided as part of the supplementary online materials

“Share upon reasonable request”

Analysis | Open Access | Published: 27 July 2021

Data sharing practices and data availability upon request differ across scientific disciplines

Leho Tedersoo ✉, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann & Tuul Sepp

Scientific Data **8**, Article number: 192 (2021) | Cite this article

19k Accesses | **66** Citations | **241** Altmetric | Metrics

https://www.nature.com/articles/s41597-021-00981-0

“Share upon reasonable request”

https://www.nature.com/articles/s41597-021-00981-0

Abstract

Data sharing is one of the cornerstones of modern science that enables large-scale analyses and reproducibility. We evaluated data availability in research articles across nine disciplines in *Nature* and *Science* magazines and recorded corresponding authors' concerns, requests and reasons for declining data sharing. Although data sharing has improved in the last decade and particularly in recent years, data availability and willingness to share data still differ greatly among disciplines. We observed that statements of data availability upon (reasonable) request are inefficient and should not be allowed by journals. To improve data sharing at the time of manuscript acceptance, researchers should be better motivated to release their data with real benefits such as recognition, or bonus points in grant and job applications. We recommend that data management costs should be covered by funding agencies; publicly available research data ought to be included in the evaluation of applications; and surveillance of data sharing should be enforced by both academic publishers and funders. These cross-discipline survey data are available from the plutoF repository.

JSDSE next steps: Data and code sharing

- Challenges: "share upon reasonable request" commonly selected
- Challenges: IRB protocols sometimes overly constrain data sharing (e.g., exempt project with deidentified data)
- Challenges: developing a process to handle genuinely sensitive information (e.g., provide synthetic data to test code)
- Challenges: T&F focus is on data sharing (not code)
- Challenges: what to do with the data and code (see Vilhuber et al, JSDSE, 2022, https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2074582)
- Undergraduate students now hired to curate data and code for JSDSE
- Will review practices in 2024

EDITORIAL

Strengthening the scientific record

H. HOLDEN THORP , VALDA VINSON , AND JAKE YESTON Authors Info & Affiliations

SCIENCE • 30 Mar 2023 • Vol 380, Issue 6640 • p. 13 • DOI: 10.1126/science.adi0333

https://www.science.org/doi/10.1126/science.adi0333

"In keeping with our commitment to reproducibility and FAIR (findable, accessible, interoperable, reusable) data principles, we require all data underlying the results in published papers to be publicly and immediately available."

Teaching reproducibility and responsible workflows: an educator and editor's perspective

Nicholas J. Horton, Amherst College

August 8, 2023, JSM, nhorton@amherst.edu

Image source: heylagostechie

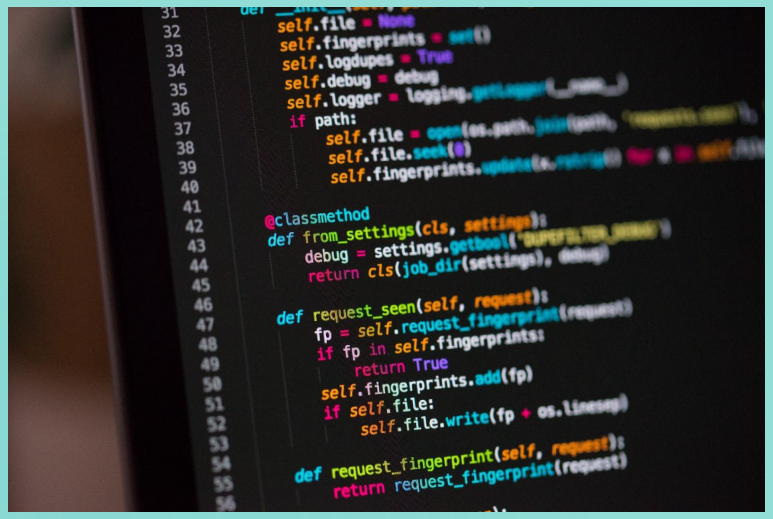


Image source: Wikicommons

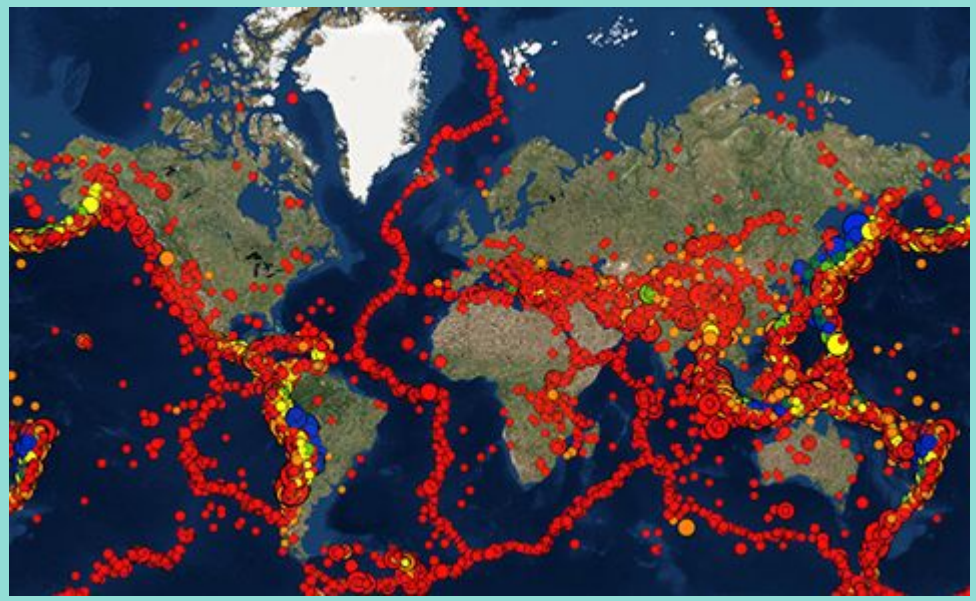


Image source: Concord Consortium

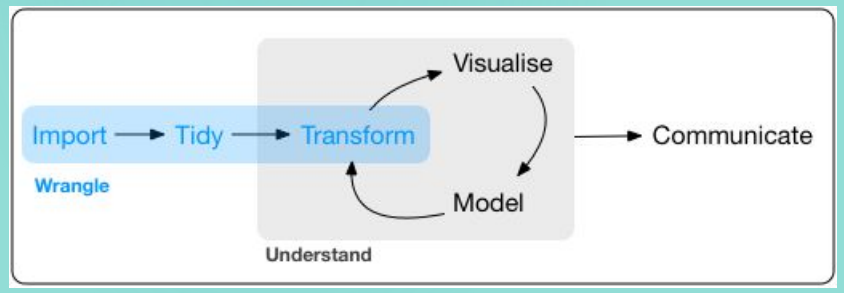


Image source: Hadley Wickham and Garrett Grolemund

thanks to NSF #1923388