

Teaching reproducibility and responsible workflows

Nicholas J. Horton, Amherst College

January, 2023, nhorton@amherst.edu

```
31 def __init__(self, path):
32     self.file = None
33     self.fingerprints = set()
34     self.logdups = True
35     self.debug = debug
36     self.logger = logging.getLogger(__name__)
37     if path:
38         self.file = open(os.path.join(path, 'reports.txt'),
39                         'a')
40         self.file.seek(0)
41         self.fingerprints.update(self._get_fingerprints())
42
43 @classmethod
44 def from_settings(cls, settings):
45     debug = settings.getbool('debug', False)
46     return cls(job_dir(settings), debug)
47
48 def request_seen(self, request):
49     fp = self.request_fingerprint(request)
50     if fp in self.fingerprints:
51         return True
52     self.fingerprints.add(fp)
53     if self.file:
54         self.file.write(fp + os.linesep)
55
56 def request_fingerprint(self, request):
57     return request_fingerprint(request)
```

Image source: Wikicommons



Image source: heylagostechie

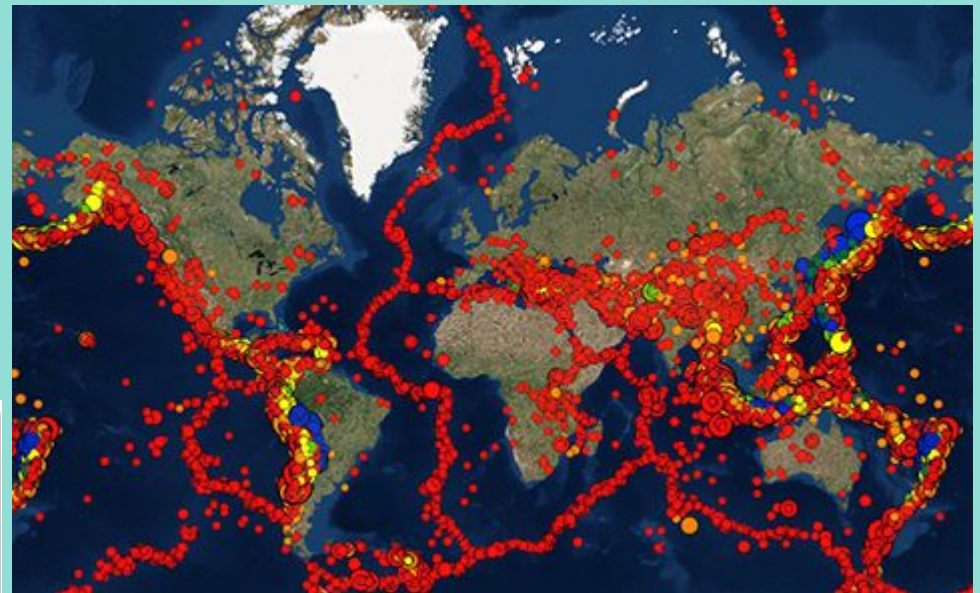


Image source: Concord Consortium

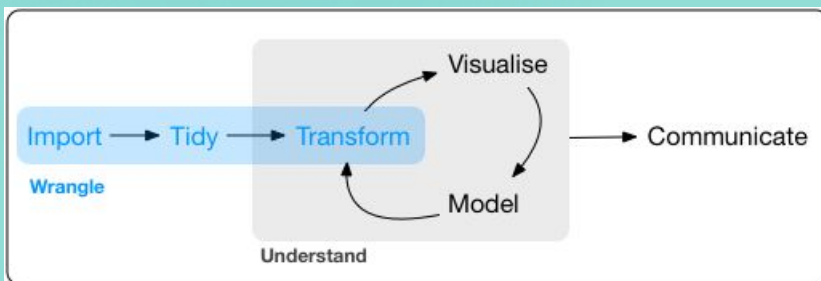


Image source: Hadley Wickham and Garrett Grolmund

Acknowledgements

- ▶ The K12 Data Tools project is joint work with Danny Pimentel and Michelle Wilkerson (commissioned paper from K-12 Data Science workshop can be found here: <https://nicholasjhorton.github.io/K12-Data-Tools>)
- ▶ The DSC-WAV project and this work is funded by the NSF (#1923388, <https://dsc-wav.github.io/www>, Ben Baumer PI)
- ▶ JSDSE issue on teaching reproducibility and responsible workflow (guest edited by Rohan Alexander, Nicholas Horton, Micaela Parker, Aneta Piekut, and Colin Rundel)
- ▶ Many key ideas derive from my collaborators: Valerie Barr, Matt Beckman, Mine Çetinkaya-Rundel, Jie Chao, Bill Finzer, Jo Hardin, Chelsey Legacy, Randall Pruim, Maria Tackett, Andy Zieffler

Plan for my talk

- ▶ Data acumen and reproducible research
Reproducibility tools (and motivation plus a look back in time)
- ▶ But what about K-12?
- ▶ Initiatives to teach reproducibility and responsible research (making change happen)
- ▶ Closing thoughts

NASEM (2019)

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science



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.

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

Mathematical concepts

Key **mathematical** 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:

- ▶ Set theory and basic logic,
- ▶ Multivariate thinking via functions and graphical displays,
- ▶ Basic probability theory and randomness,
- ▶ Matrices and basic linear algebra,
- ▶ Networks and graph theory, and
- ▶ Optimization.

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**, Idea of “computational essay”: more later
- ▶ Programming concepts,
- ▶ Data structures, and
- ▶ **Simulations.**

Statistical concepts

Important **statistical foundations** might include the following:

- ▶ Variability, uncertainty, sampling error, and inference;
- ▶ Multivariate thinking;
- ▶ Nonsampling error, design, experiments (e.g., A/B testing), biases, confounding, and causal inference;
- ▶ Exploratory data analysis;
- ▶ Statistical modeling and model assessment; and
- ▶ Simulations and experiments

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”
as part of modeling cycle:
more later!

- ▶ 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 visualization concepts

Key **data description and visualization** 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 consistency checking,
- Exploratory data analysis,
- Grammar of graphics,
- Attractive and sound static visualizations,
- Dynamic visualizations and dashboards.

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.

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.**

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.

What is reproducibility? Modern Data Science with R (appendix D)

<https://mdsr-book.github.io/mdsr2e/>

To further explicate the distinction between *replicability* and *reproducibility*, recall that scientists are legendary keepers of lab notebooks. These notebooks are intended to contain all of the information needed to carry out the study again (i.e., replicate): reagents and other supplies, equipment, experimental material, etc. Modern software tools enable scientists to carry this same ethos to data analysis: Everything needed to repeat the analysis (i.e., reproduce) should be recorded in one place.

Basic tools for reproducible analysis and responsible workflow

- ▶ “Reproducibility is the ultimate standard by which scientific findings are judged” Dynamic Documents (Xie, 2015)
- ▶ Idea of “literate programming” (Knuth, 1984): a single file is knit/woven/rendered to generate a program and documentation
- ▶ Sweave adaptation (Leisch, 2002): create a report that incorporates graphics, other output and text from a single file
- ▶ Facilitates a “computational essay” (Fleischer et al, ICOTS II, Hüsing and Podworny, IASE2021) that fully describes an analysis
- ▶ Avoids the perils of cut and paste!
- ▶ Necessary (if not sufficient) component of a reproducible workflow?

Example of reproducibility tool: RMarkdown/quarto

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar below the menu bar contains icons for file operations and a 'Go to file/function' search bar. The R version is 4.2.1.

The main editor window shows a file named 'fishdata_just_R.Rmd'. The source editor displays the following R Markdown code:

```
1 |---
2 |title: "K12 data tools vignette in R: where are the lobsters?"
3 |author: "Nicholas Horton (nhorton@amherst.edu), Danny Pimental,
4 |and Michelle Wilkerson"
5 |date: "August 29, 2022"
6 |output:
7 |  pdf_document:
8 |    fig_height: 7
9 |    fig_width: 8
10 |  toc: true
```

The environment pane on the right shows the 'Global Environment' is empty. The file explorer on the bottom right shows the project files:

Name	Size	Modified
..		
.Rhistory	0 B	Nov 6, 2022, 7:3
fishdata_just_R.pdf	6.5 MB	Nov 6, 2022, 7:4
fishdata_just_R.Rmd	6.7 KB	Nov 6, 2022, 7:3
fishdata.csv	87.1 KB	Nov 6, 2022, 7:3
project.Rproj	205 B	Nov 6, 2022, 7:4

The console at the bottom shows the R version and copyright information:

```
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

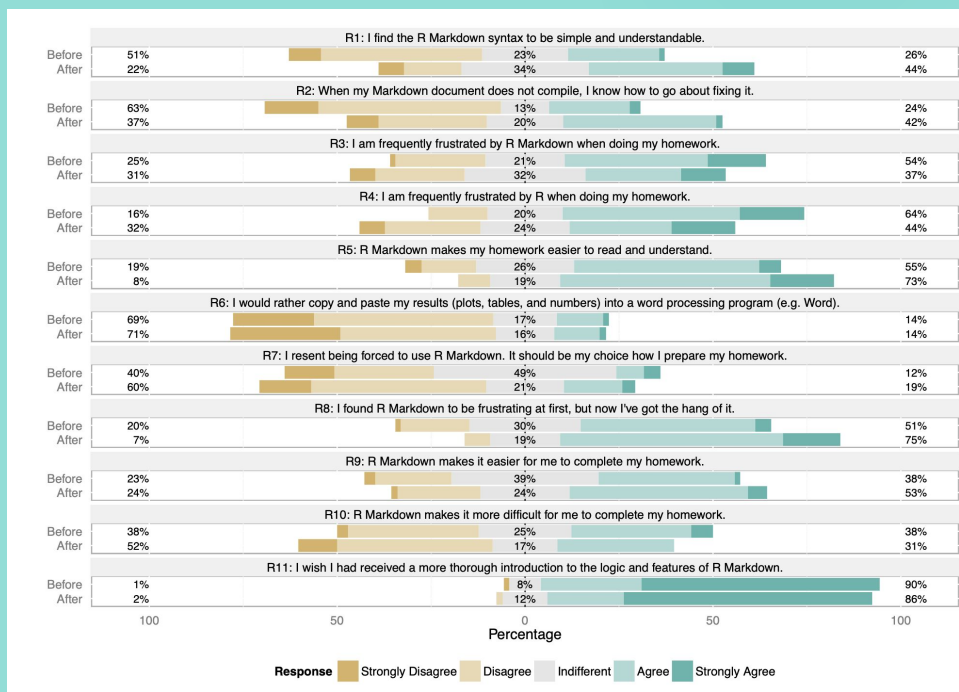
Try it at <https://posit.cloud/content/4896839>

Baumer et al (TISE, 2014)

R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics

<https://escholarship.org/uc/item/90b2f5xh>

Includes reports of student experiences from Duke University and Smith College



Baumer et al (TISE, 2014)

R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics

<https://escholarship.org/uc/item/90b2f5xh>

Student experiences from Duke University and Smith College

“In our experience at two very different institutions with very different student bodies, RMarkdown made a welcomed improvement to the traditional copy-and-paste workflow.”

Quarto (R/Python/Julia)

Welcome to Quarto

Quarto® is an open-source scientific and technical publishing system built on [Pandoc](#)

- Create dynamic content with [Python](#), [R](#), [Julia](#), and [Observable](#).
- Author documents as plain text markdown or [Jupyter](#) notebooks.
- Publish high-quality articles, reports, presentations, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
- Author with scientific markdown, including equations, citations, crossrefs, figure panels, callouts, advanced layout, and more.

Get Started

Guide

Try it at <https://quarto.org>

Quarto (Julia): successor to RMarkdown

```
---  
title: "Plots Demo"  
author: "Norah Jones"  
date: "5/22/2021"  
format:  
  html:  
    code-fold: true  
jupyter: julia-1.8  
---
```

Parametric Plots

Plot function pair $(x(u), y(u))$.
See @fig-parametric for an example.

```
``{julia}  
#| label: fig-parametric  
#| fig-cap: "Parametric Plots"
```

using Plots

```
plot(sin,  
      x->sin(2x),  
      0,  
      2π,  
      leg=false,  
      fill=(0,:lavender))  
...
```

Plots Demo

AUTHOR
Norah Jones

PUBLISHED
May 22, 2021

Parametric Plots

Plot function pair $(x(u), y(u))$. See [Figure 1](#) for an example.

► Code

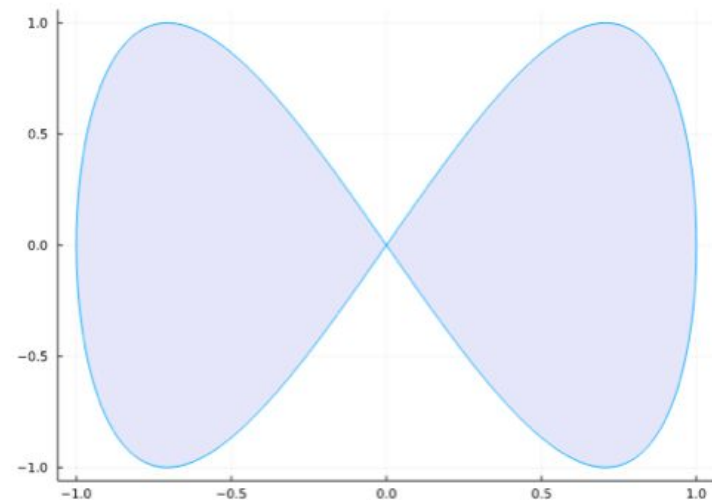


Figure 1: Parametric Plots

Another tool: version control

- ▶ 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

	COMMENT	DATE
○	CREATED MAIN LOOP & TIMING CONTROL	14 HOURS AGO
○	ENABLED CONFIG FILE PARSING	9 HOURS AGO
○	MISC BUGFIXES	5 HOURS AGO
○	CODE ADDITIONS/EDITS	4 HOURS AGO
○	MORE CODE	4 HOURS AGO
○	HERE HAVE CODE	4 HOURS AGO
○	AAAAA	3 HOURS AGO
○	ADKFJSLKDFJSDKLFJ	3 HOURS AGO
○	MY HANDS ARE TYPING WORDS	2 HOURS AGO
○	HAAAAAAAAAANDS	2 HOURS AGO

AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES GET LESS AND LESS INFORMATIVE.

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.”

Jenny Bryan (Happy Git with R)

Exposure: If someone needs to see your work or if you want them to try out your code, they can easily get it from GitHub. If they use Git, they can clone or fork your repository. If they don't use Git, they can still browse your project on GitHub like a normal website and even grab everything by downloading a zip archive.

Be a keener! If you care deeply about someone else's project, such as an R package you use heavily, you can track its development on GitHub. You can watch the repository to get notified of major activity. You can fork it to keep your own copy. You can modify your fork to add features or fix bugs and send them back to the owner as a proposed change.

Collaboration: If you need to collaborate on data analysis or code development, then everyone should use Git. Use GitHub as your clearinghouse: individuals work independently, then send work back to GitHub for reconciliation and transmission to the rest of the team. The advantage of Git/GitHub is highlighted by comparing these two ways of collaborating on a document:



ReScience Editions

@ReScienceEds



Ten Years [#reproducibility #challenge](#).
rescience.github.io/ten-years/

Did you ever try to run an old code that you wrote for a scientific article you published years ago? Did you encounter any problems? Were you successful? We are curious to hear your story.

TEN YEARS REPRODUCIBILITY CHALLENGE

R E S C I E N C E S P E C I A L I S S U E
F R E E T O R E A D - F R E E T O P U B L I S H

Ten Years Reproducibility Challenge

Did you ever try to run old code that you wrote for a scientific article you published years ago? Did you encounter any problems? Were you successful? We are curious to hear your story. This is the reason why we are editing a special issue of ReScience to collect these stories.

The ten years reproducibility challenge is an invitation for researchers to try to run the code they've created for a scientific publication that was published more than **ten years ago**. This code can be anything (statistical analysis, numerical simulation, data processing, etc.), can be written in any language and can address any scientific domain. The only mandatory condition to enter the challenge is to have published a scientific article **before 2010**, in a journal or a conference with proceedings, which contains results produced by code, irrespectively of whether this code was published in some form at the time or not.

[Addict Behav.](#) Author manuscript; available in PMC 2014 Oct 1.

PMCID: PMC3725189

Published in final edited form as:

NIHMSID: NIHMS483578

[Addict Behav. 2013 Oct; 38\(10\): 2532–2540.](#)

PMID: [23778317](#)

Published online 2013 May 21. doi: [10.1016/j.addbeh.2013.04.009](#)

Characterizing High School Students Who Play Drinking Games Using Latent Class Analysis

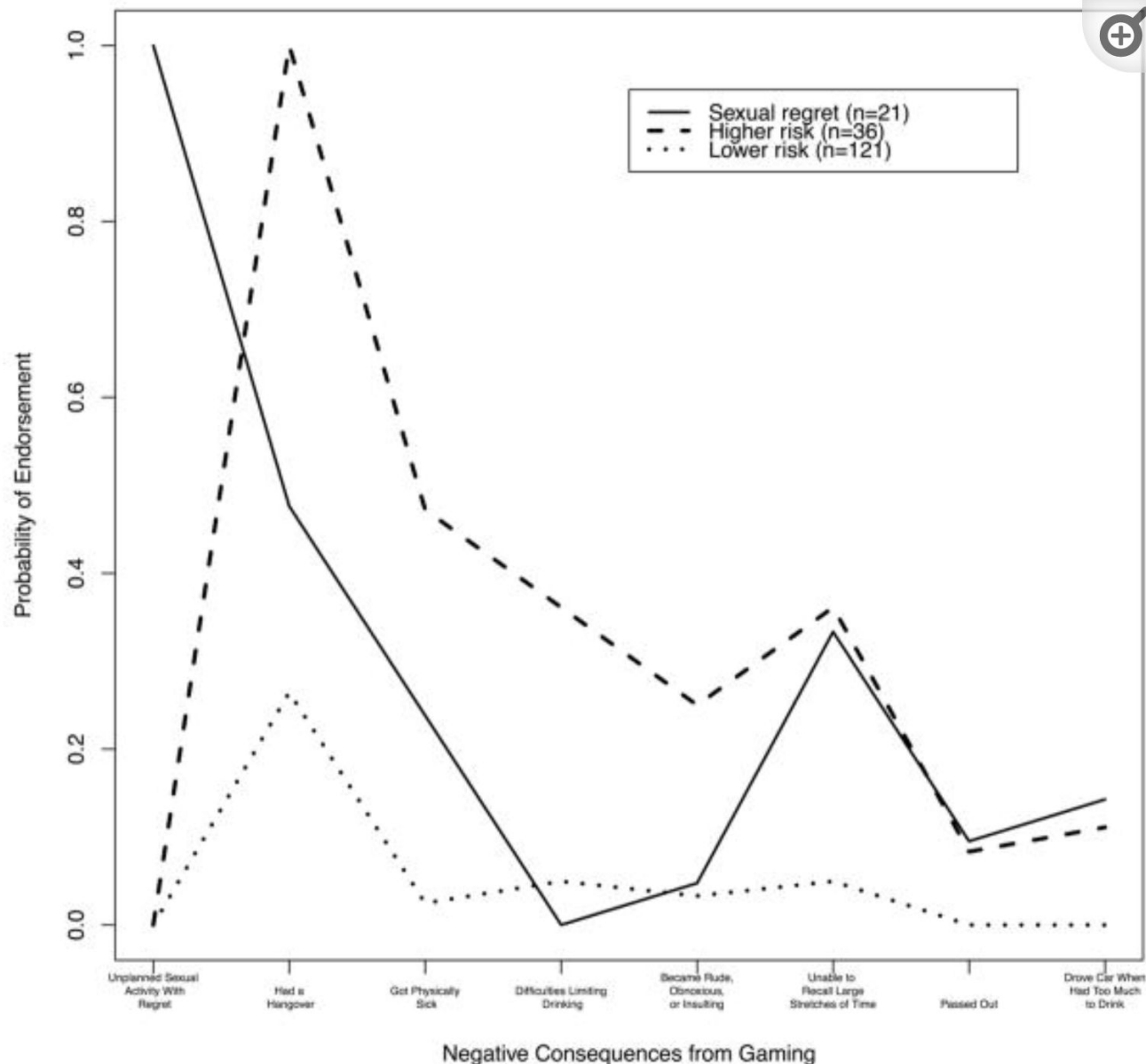
[Brian Borsari](#),¹ [Byron L. Zamboanga](#),² [Christopher Correia](#),⁴ [Janine V. Olthuis](#),⁴ [Kathryne Van Tyne](#),⁵
[Zoe Zadworny](#),² [Joel R. Grossbard](#),⁶ and [Nicholas J. Horton](#)²

► [Author information](#) ► [Copyright and License information](#) [Disclaimer](#)

Demographics, Risky Drinking, Gaming Specific Behaviors, Game-Related Consequences, and Alcohol-Related Cognitions for Total Gaming Sample and Three Classes of Gamers

Variable	% /Mean	<i>n</i>	Class 2			<i>p</i> -value/ Effect Size
			Class 1 “Lower-Risk” Gamers (<i>n</i> =121, 68%)	“Higher-risk” Gamers (<i>n</i> =36, 20%)	Class 3 “Sexual Regret” Gamers (<i>n</i> = 21, 12%)	
Demographics						
Age (Mean, Median)	16.3, 16	178	16	17	17	<i>p</i> = 0.085
Male *	49%	178	43%	56%	76%	<i>p</i> = 0.014/ φ _c =.22
Typical Grade	A’s and B’s	176	A’s and B’s	Mostly B’s	Mostly B’s/B’s&C’s	<i>p</i> = 0.080
Varsity Sport Participation	65%	171	64%	78%	53%	<i>p</i> = 0.141/ φ _c =.15
Alcohol Initiation ≥ 14 Years *	61%	175	68%	46%	45%	<i>p</i> = 0.017/ φ _c =.22

Zamboanga et al (2013) replication



Zamboanga et al (2013) replication

Full disclosure: my code from 2012 was clunky!

But the good news is that it was mostly reproducible 10 years later

- ▶ used Sweave (a precursor to RMarkdown, still works) [+]
- ▶ clear provenance for datasets [+]
- ▶ multiple analyses dated (to clarify which is which), but not kept within a version control system [-]
- ▶ syntax and style not ideal (but not terrible)

More on replication tools

- ▶ Reproducibility and responsible workflow provide a necessary foundation for data science and statistics practice
- ▶ We are seeing uptake in university and graduate programs (more later)
- ▶ But what about schools?
- ▶ When and how should we teach these topics?
- ▶ What tools should we use?

Growth of K12 Data Science

- ▶ In a world defined by data, we can't wait to introduce students in K-12 to the opportunities (and challenges) in making sense of it
- ▶ Increasing growth of K12 Data Science:
 - ▶ NASEM workshop,
<https://www.nationalacademies.org/our-work/foundations-of-data-science-for-students-in-grades-k-12-a-workshop>
 - ▶ GAISE College report and GAISE II,
[https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-\(gaise\)-reports](https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports)

Revised Guidelines for Assessment and Instruction in Statistics Education (GAISE) College report (2016)

- ▶ Teach statistical thinking.
- ▶ **Teach statistics as an investigative process of problem-solving and decision-making.**
- ▶ Give students experience with multivariable thinking.
- ▶ Focus on conceptual understanding.
- ▶ Integrate real data with a context and purpose.
- ▶ Foster active learning.
- ▶ **Use technology to explore concepts and analyze data.**
- ▶ Use assessments to improve and evaluate student learning.

[https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-\(gaise\)-reports](https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports)

Revised K12 GAISE Guidelines

Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)

A Framework for Statistics and Data Science Education

Anna Bargagliotti (co-chair)
Christine Franklin (co-chair)
Pip Arnold
Rob Gould
Sheri Johnson
Leticia Perez
Denise A. Spangler

Original K-12 report written in 2005, published in 2007,
revised (and renamed “GAISE II”) in 2020

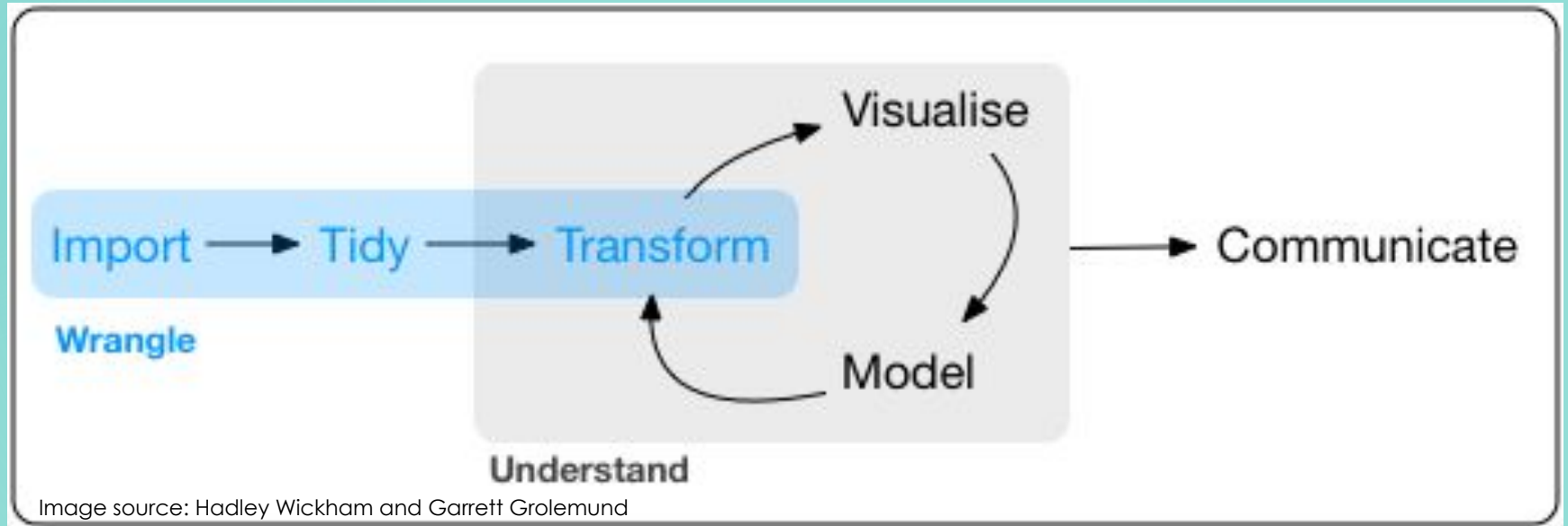
Revised Guidelines for Assessment and Instruction in Statistics Education PreK-12 [GAISE II] report (2020)

- ▶ **Importance of questioning through the problem-solving cycle (see Lee et al, SERJ, 2022)**
- ▶ Importance of design and considering different data types
- ▶ Inclusion of multivariate thinking
- ▶ Role of probabilistic thinking
- ▶ **Shifts and deepening of technology**
- ▶ Importance of communication

[https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-\(gaise\)-reports](https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports)

Problem-solving cycle

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



Data Tools for Data Science

- ▶ Lots of data science now taking place in K12 (with **much** more to come, see for example <https://doe.virginia.gov/boe/meetings/2022/04-apr/item-g.pdf>)
- ▶ Good news: there are many tools for doing and teaching data analysis and data science
- ▶ What is the state of the art for reproducibility?

Data Tools for K12 Data Science

► What are we hoping that students will learn?



BOOTSTRAP
Equity • Scale • Rigor



Bootstrap:Data Science

Evidence-based, integrated materials for grade 7-12 Social Studies classes

- Leverage students' curiosity about the world around them to inspire real data analysis and original research.
- Lessons are available for data visualization, measures of center and spread, programming, linear regression, and more.
- Mix and match to create anything from a [one-week intro to a full-year course!](#)



Common Online Data Analysis Platform (CODAP)

Open-source software for dynamic data exploration

For Educators

For Developers

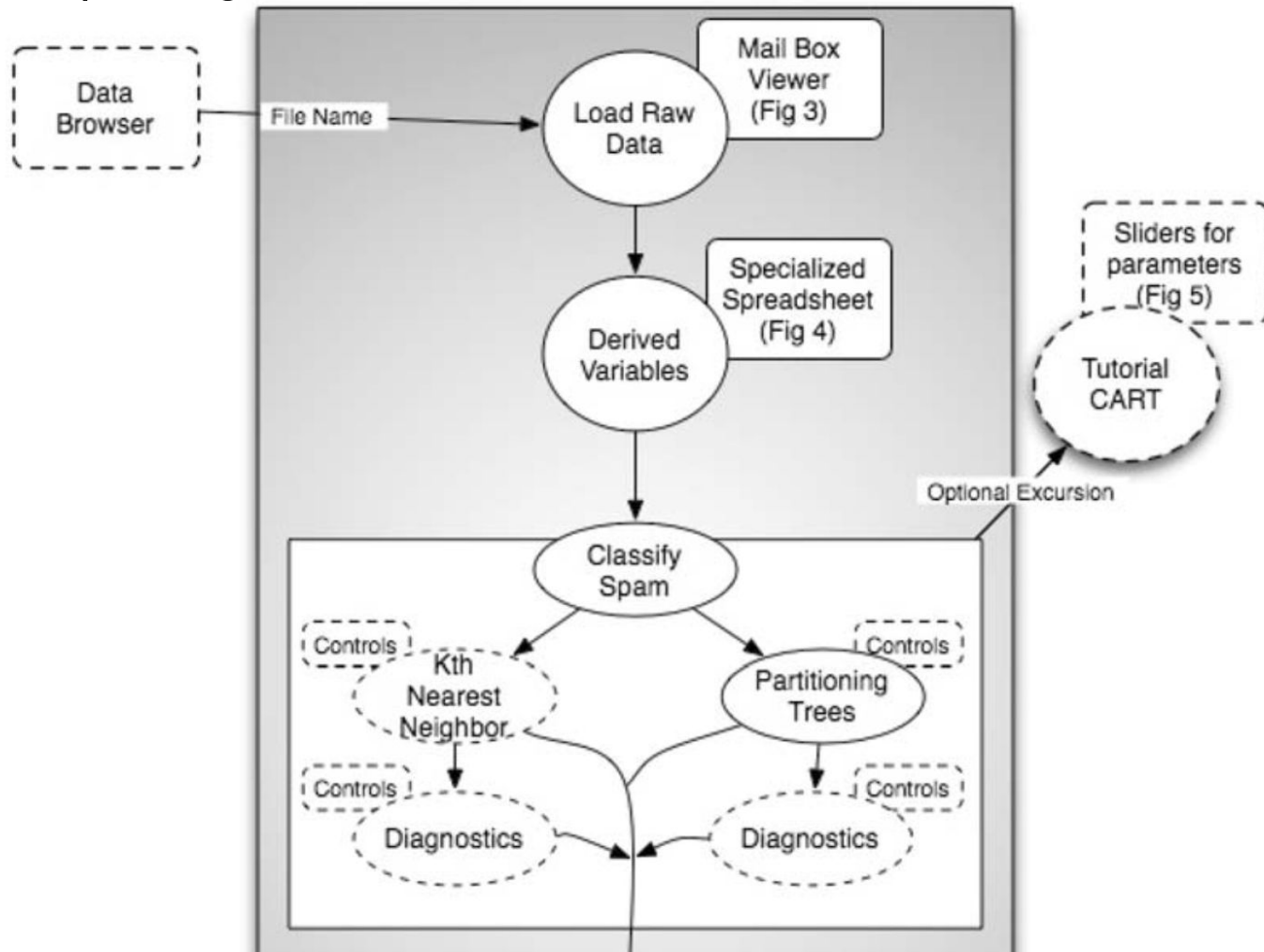


Vision for dynamic documents

Dynamic, Interactive Documents for Teaching Statistical Practice

303

Nolan and Temple Lang, ISR, 2007, <https://doi.org/10.1111/j.1751-5823.2007.00025.x>



Prior work on data tools

- ▶ Biehler (1997, ISR) “Software for Learning and for Doing Statistics”,
<https://doi.org/10.1111/j.1751-5823.1997.tb00399.x>
- ▶ McNamara (2019, TAS) “Key attributes of a modern statistical computing tool”,
<https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1482784>, see also <https://arxiv.org/abs/1610.00984>
- ▶ We adapted the framework of McNamara to account for considerations specific to K12 education

Framework (based on McNamara, 2019)

- ▶ **Accessibility:** Includes cost, simplicity of cloud-based tools, disability access, multilingual support
- ▶ **Ease of entry:** Clarity about how the tool works; includes consideration of students' conceptions of data and developmental appropriateness
- ▶ **Data as a first-order object:** Data as primary interest: hierarchical vs. tabular formats, viewing data; key to building “students conception of data”

Framework (based on McNamara, 2019)

- ▶ **Data analysis cycle and reproducible workflows:** Iterative cycle of posing questions, exploring data, visualizing results, modeling, model assessment, and communicating results; reproducing data wrangling, analyses, and explorations
- ▶ **Interactivity:** Support for direct interaction with data, e.g., pinch, click-and-drag, brushing, hovering
- ▶ **Flexible plot creation:** Univariate, bivariate, and multivariate displays with ability to augment graphics in a variety of ways

most relevant
today

Framework (based on McNamara, 2019)

- ▶ **Inferential analysis:** Reasoning with samples and inferring beyond data; support for simulations and resampling; offering probabilistic or uncertain expressions of data
- ▶ **Non-standard data:** Working with multiple forms of data such as spatial data, network data, etc.
- ▶ **Extensibility:** included in prior frameworks: important for the future but beyond our scope

Genres of Data Tools

- ▶ Spreadsheets
 - ▶ Google Sheets
 - ▶ Excel
- ▶ Visual tools
 - ▶ CODAP
 - ▶ iNZight
 - ▶ Tuva
 - ▶ Tableau
- ▶ Scripting languages
 - ▶ Python
 - ▶ R
 - ▶ Julia

Illustrative example: lobsters

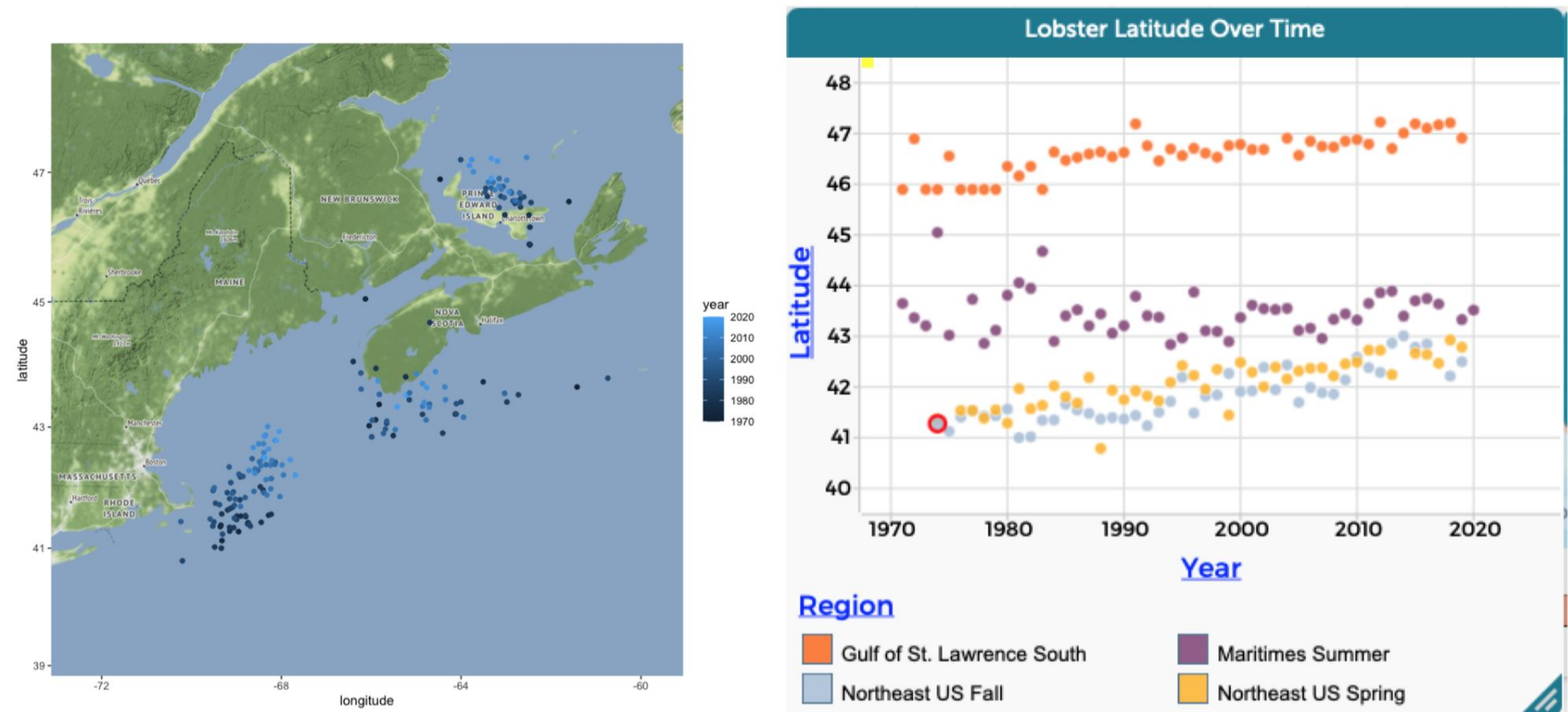


Figure 1: Left: Map of groups of lobster between 1970 and 2020, colored by year. Right: Scatterplot of latitude of lobster populations over time, colored by region.

Homarus americanus (lobsters)

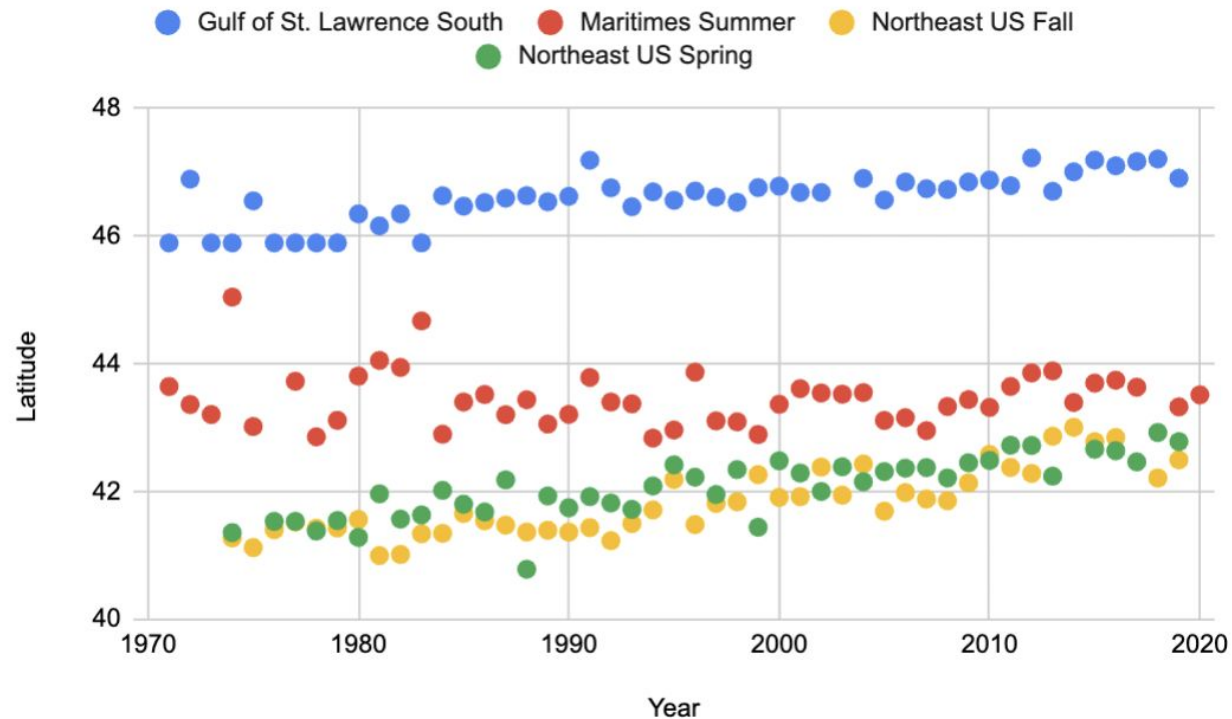
- ▶ Data on mean location of lobsters from 1970 through 2020 in various regressions off the northeast of the United States and Canada
- ▶ There is evidence that the populations have been moving northward over time
- ▶ Exploring such movement is often included as an activity in math or science class
- ▶ Includes spatial data (maps) and time series

Data, code, illustrated examples, and interactive links available at:
<https://nicholasjhorton.github.io/K12-Data-Tools/dsd.html>

Lobsters in spreadsheets

G3 fx =IF(F3="Gulf of St. Lawrence South", D3, "")

	A	B	C	D	E	F	G	H	I	J	
1	Common name	Latin name	Year	Latitude	Longitude	Region	Gulf of St. Lawrence	Maritimes Summer	Northeast US Fall	Northeast US Spring	
2	American lobster	Homarus americanus	1970			Maritimes Summer					
3	American lobster	Homarus americanus	1971	45.8967312	-62.476134	Gulf of St. Lawrence	45.8967312				
4	American lobster	Homarus americanus	1971	43.647823	-61.419999	Maritimes Summer		43.647823			
5	American lobster	Homarus americanus	1972	46.8950897	-64.468539	Gulf of St. Lawrence	46.8950897				
6	American lobster	Homarus americanus	1972	43.36638	-65.825705	Maritimes Summer		43.36638			
7	American lobster	Homarus americanus	1973	45.8967312	-62.476134	Gulf of St. Lawrence	45.8967312				
8	American lobster	Homarus americanus	1973								
9	American lobster	Homarus americanus	1974								
10	American lobster	Homarus americanus	1974								
11	American lobster	Homarus americanus	1974								
12	American lobster	Homarus americanus	1974								



Thoughts about spreadsheets

- ▶ Commonly accessible (Excel or Google Sheets)
- ▶ Often used for simpler analyses, ease of entry
- ▶ Offer rudimentary graphics and tables
- ▶ Data at the fore (easy to review and examine individual points)
- ▶ Challenging to undertake some “data moves” (Erickson et al, TISE, <https://escholarship.org/uc/item/0mg8m7g6>)
- ▶ Challenging to undertake multivariate visualization
- ▶ actively “**prevent**” **reproducibility** (see Biehler 1997 and McNamara 2019)

Lobsters in CODAP

Changes in Fish Habitat CODAP Extension

Tables Graph Map Slider Calc Text Plugins

0.85 Story Builder

1 The American Lobster 2 Moving North 3 Graphing the Changes 4 Looking at More Populations 5 West Coast Rainbow Star

The American Lobster

In recent years, fishers have noticed that American Lobster populations are moving. Fishers in Rhode Island and Boston are finding it harder to catch lobsters, while fishers in Maine and Canada are finding it easier.


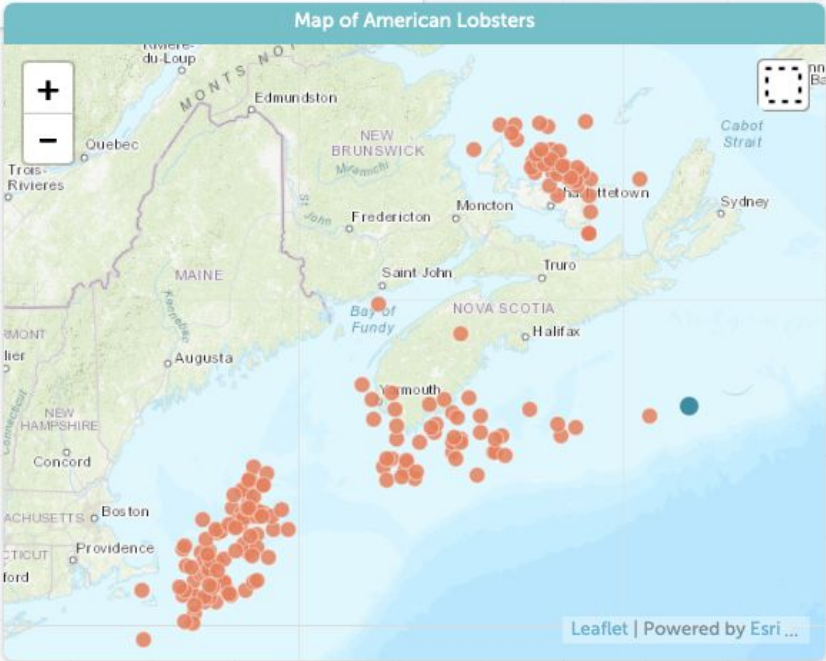


Photo courtesy of Gulf of Maine Research Institute

Look at the map on the right.

- What do you notice?
- What do you think the dots represent?
- Why are the dots clustered in this way?

Map of American Lobsters



Leaflet | Powered by Esri ...

Try it at the following link: <https://tinyurl.com/dsd-codap>

Thoughts about visual tools

- ▶ Excellent at accessibility and “data as a first-order, persistent object”
- ▶ Facilitate interactive exploration and flexible plot creation
- ▶ Support for inference somewhat limited
- ▶ **Minimal support for reproducibility** (moves made are not recorded)

Lobsters in R (RMarkdown)

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The toolbar contains icons for file operations and a search bar. The main editor window shows the R Markdown document 'fishdata_just_R.Rmd' with the following content:

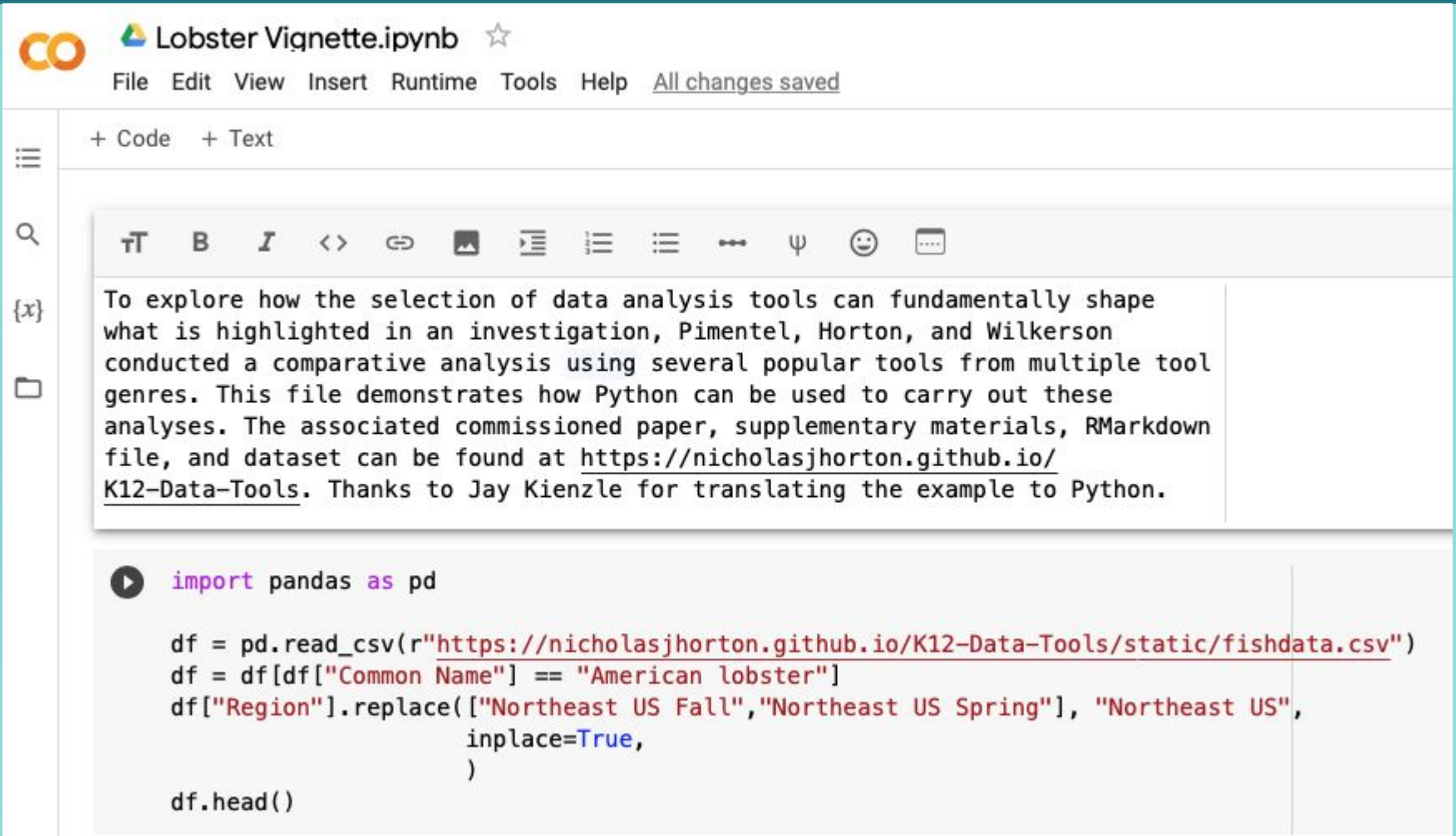
```
1 |---
2 |title: "K12 data tools vignette in R: where are the lobsters?"
3 |author: "Nicholas Horton (nhorton@amherst.edu), Danny Pimental,
4 |and Michelle Wilkerson"
5 |date: "August 29, 2022"
6 |output:
7 |  pdf_document:
8 |    fig_height: 7
9 |    fig_width: 8
10 |  toc: true
```

The console at the bottom left shows the R version 4.2.1 (2022-06-23) and copyright information. The file explorer on the right shows the 'project' directory with the following files:

Name	Size	Modified
..		
.Rhistory	0 B	Nov 6, 2022, 7:3
fishdata_just_R.pdf	6.5 MB	Nov 6, 2022, 7:4
fishdata_just_R.Rmd	6.7 KB	Nov 6, 2022, 7:3
fishdata.csv	87.1 KB	Nov 6, 2022, 7:3
project.Rproj	205 B	Nov 6, 2022, 7:4

Try it at <https://posit.cloud/content/4896839>

Lobsters in Python (JupyterHub)



The screenshot shows a Jupyter Notebook interface. At the top, the title bar reads "Lobster Vignette.ipynb" with a star icon. Below the title bar is a menu bar with options: File, Edit, View, Insert, Runtime, Tools, Help, and a link "All changes saved". The left sidebar contains icons for a menu, search, and a file explorer. The main area is divided into two sections. The top section is a text cell containing a paragraph about data analysis tools. The bottom section is a code cell containing Python code for loading and filtering a CSV file.

File Edit View Insert Runtime Tools Help [All changes saved](#)

+ Code + Text

To explore how the selection of data analysis tools can fundamentally shape what is highlighted in an investigation, Pimentel, Horton, and Wilkerson conducted a comparative analysis using several popular tools from multiple tool genres. This file demonstrates how Python can be used to carry out these analyses. The associated commissioned paper, supplementary materials, RMarkdown file, and dataset can be found at <https://nicholasjhorton.github.io/K12-Data-Tools>. Thanks to Jay Kienzle for translating the example to Python.

```
import pandas as pd

df = pd.read_csv(r"https://nicholasjhorton.github.io/K12-Data-Tools/static/fishdata.csv")
df = df[df["Common Name"] == "American lobster"]
df["Region"].replace(["Northeast US Fall", "Northeast US Spring"], "Northeast US",
                    inplace=True,
                    )

df.head()
```

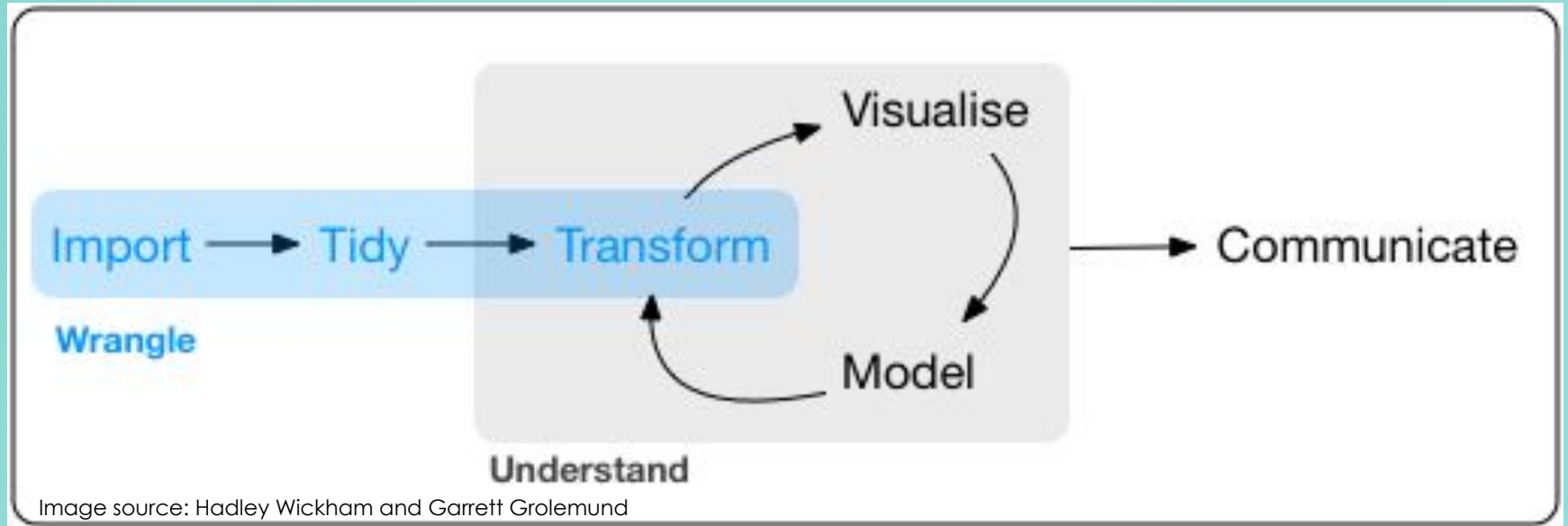
Try it at https://colab.research.google.com/drive/1teSZfFBm_o_2oe0-AMWhokRFGqV40tgp?usp=sharing

Thoughts about coding/scripting tools

- ▶ Python, R, Julia, Pyret are freely available
- ▶ Excellent graphical displays
- ▶ Most functionality (these are professional tools not designed for teaching [Pyret an exception])
- ▶ Tradeoff/tension: steep learning curve,
- ▶ Challenge for both students and instructors
- ▶ Potential for simplified interfaces (e.g., Data8 for Python, Pyret, Project MOSAIC for R)
- ▶ **best tools in terms of support for reproducibility**

Problem-solving cycle

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



Comparisons of genres

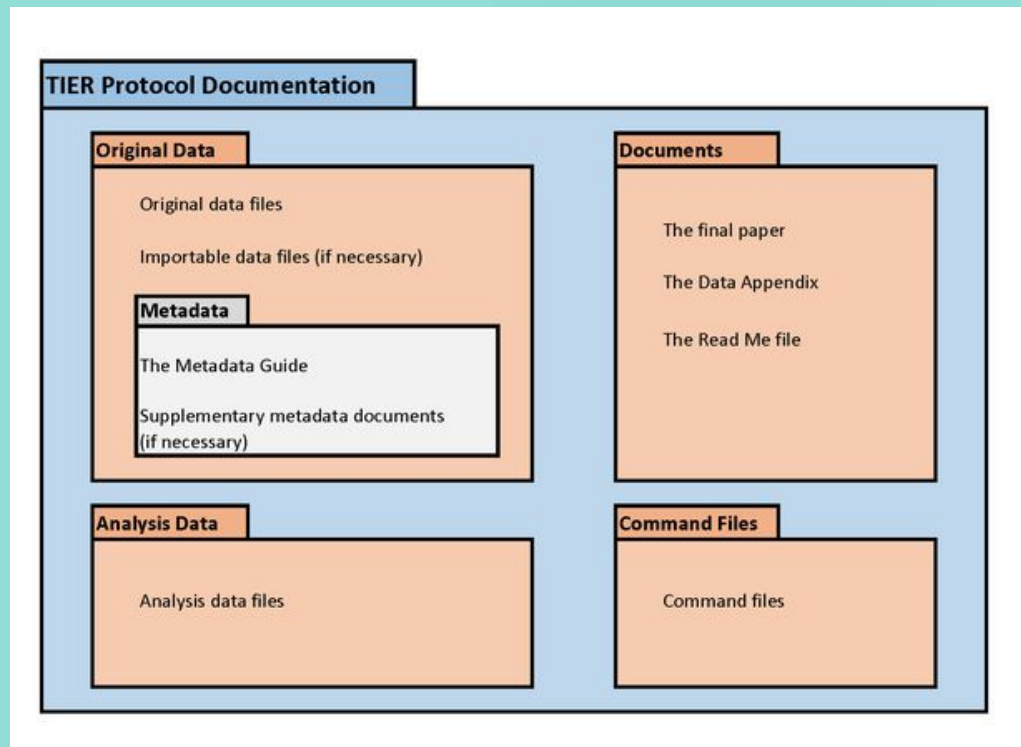
- ▶ Need to develop a **learning progression** that involves repeated opportunities to practice the entire data analysis cycle
- ▶ This will likely involve **multiple** tools **introduced** and **reinforced** over time (see Biehler, 2018 and commissioned paper)
- ▶ Biehler (ISR, 1997) challenged the community to improve tools for teaching: much more work is still needed, particularly as we think about **fostering reproducibility**

Some questions

- ▶ Why are these methods important for students to learn?
- ▶ What are the capacities we want them to develop?
- ▶ What are the skills that we need them to master to fluently utilize these methods?
- ▶ What are best practices for teaching these methods?
- ▶ Where can these methods be incorporated into the K-12 and college curricula?
- ▶ How do we assess how effectively students can apply these methods?

Foundational work: TIER protocol 3.0

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

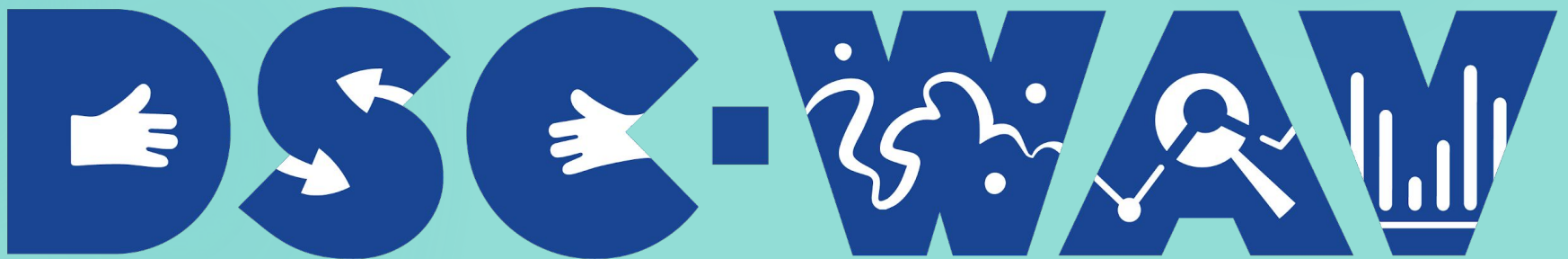


DSC-WAV (Wrangle-Analyze-Visualize)

- ▶ NSF funded effort from the Harnessing the Data Revolution (HDR) Data Science Corps (DSC) initiative:

<https://dsc-wav.github.io/www>

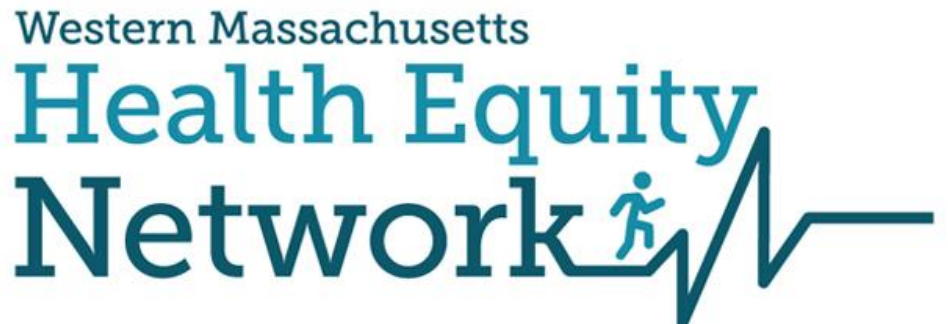
DATA SCIENCE CORPS



WRANGLE•ANALYZE•VISUALIZE

DSC-WAV (Wrangle-Analyze-Visualize)

- ▶ <https://dsc-wav.github.io/www>
- ▶ Undergraduate students working on Data Science for Social Good projects



girls
inc.

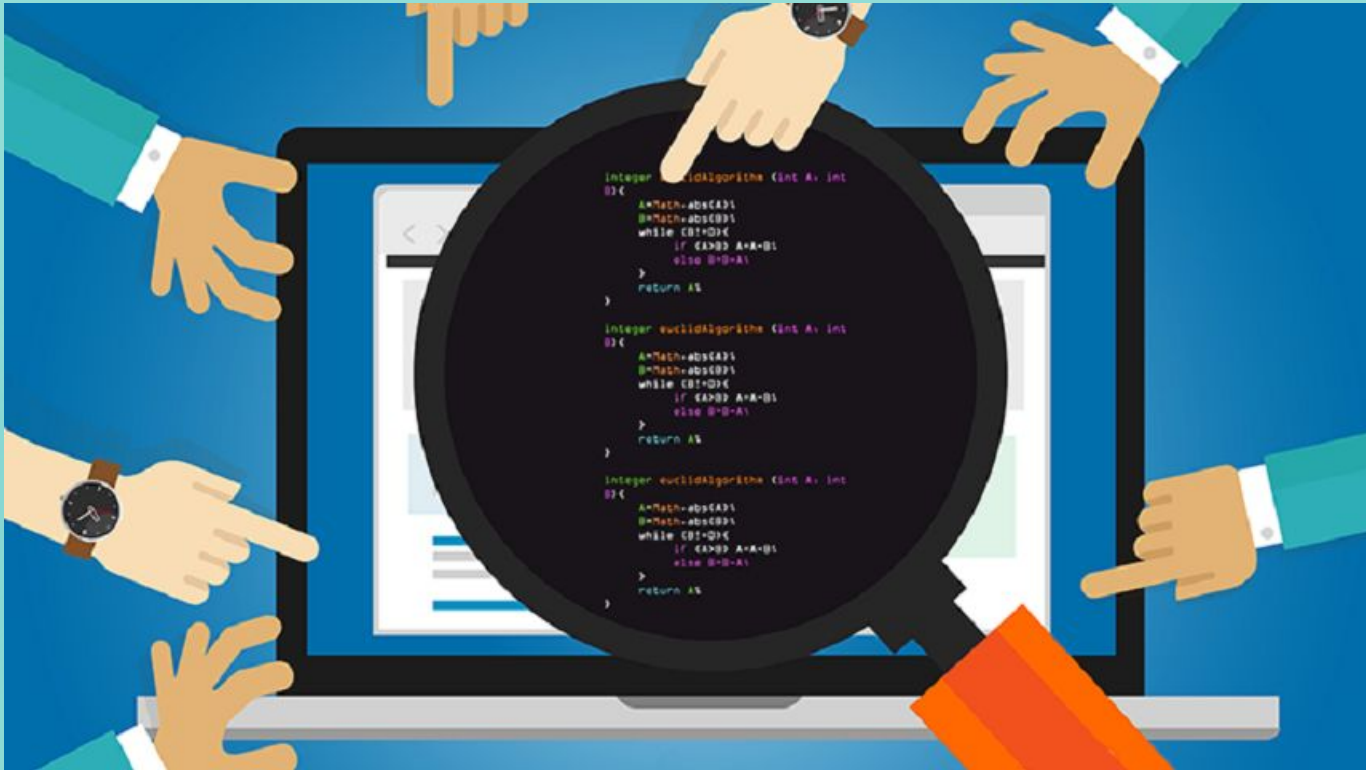
of the Valley



The Nature
Conservancy
Protecting nature. Preserving life.™



Agile and scrum for undergraduates: workflow in action



Source: techgig.com

Agile and scrum for undergraduates: workflow in action

► <https://hdsr.mitpress.mit.edu/pub/nvflcexe/release/1>

“While many of these courses and programs teach students relevant data science skills, we can expect coursework to develop students’ data acumen only so far. It is unclear whether coursework alone is enough to provide students with the experiences with data and computing they need to be successful in tomorrow’s workplace.”



Agile and scrum for undergraduates: workflow in action

The work on the project is organized into a series of short **sprints** to break up large tasks.

- ▶ Subtasks are organized into a **backlog** to identify priorities for that stage of the analysis.
- ▶ The team and stakeholders (faculty and community organization liaison) meet regularly (**standups**) to share results and make adjustments in advance of the next sprint.
- ▶ **Kanban** project boards, implemented using Trello or GitHub Projects, are used to review the backlog and team progress.
- ▶ **Code review**, implemented using GitHub pull requests, is included as a regular part of the process.

Agile and scrum for undergraduates: workflow in action

The work on the project is organized into a series of short **sprints** to break up large tasks.

- ▶ Sprint **demos** are places where current results are presented and discussed in the context of the broader goals of the project.
- ▶ Sprint **retrospectives** are used to identify issues with the process and ways that the team might improve their work.

Agile and scrum for undergraduates: workflow in action

- “Facilitating team-based data science: Lessons learned from the DSC-WAV project”, *Foundations of Data Science* (Legacy et al, <https://www.aims sciences.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."



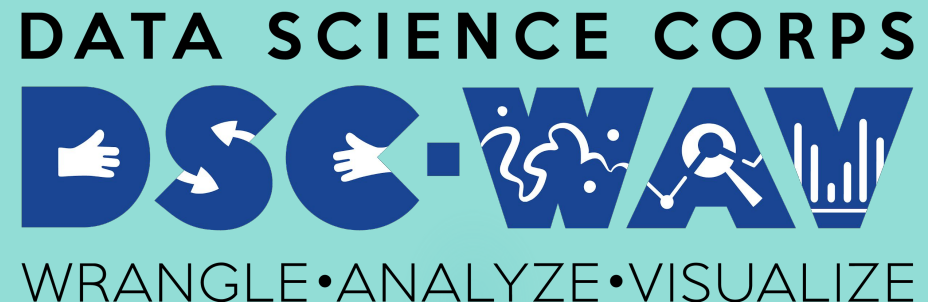
Source: smartbear.com



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

DSC-WAV Lessons Learned

- ▶ Many challenges to helping undergraduate students develop the ability to “think with data”
- ▶ Our courses and programs need to adapt to give them necessary workforce skills as analysts
- ▶ DSC-WAV projects have provided a starting point but more reinforcement is needed
- ▶ Lots of work needed to scale out programs at two- and four-year schools



Toronto Reproducibility Workshop

A two-day workshop focusing on reproducibility in data-centric analysis

<https://rohanalexander.com/reproducibility.html>

Toronto Data Workshop on Reproducibility

Posted by [Lauren Kennedy](#) on 8 February 2021, 10:33 pm

I (Lauren not Andrew writing) will be speaking at an upcoming online workshop on reproducibility (free and open). [More details here](#). Looking at the talk outlines, I'm really looking forward to it. I think we can generally agree that reproducibility is a good thing, and something we want to strive for, but in practice there's a lot of complexity to a real world reproducibility workflow. I'm by no means an expert, so I'm hoping to pick up so new tips, tricks and reproducible perspectives!

The Faculty of Information and the Department of Statistical Sciences at the University of Toronto are excited to host a two-day conference bringing together academic and industry participants on the critical issue of reproducibility in applied statistics and related areas. The conference is free and will be hosted online on Thursday and Friday 25–26 February 2021. Everyone is welcome, you don't need to be affiliated with a university, and you can register [here](#).

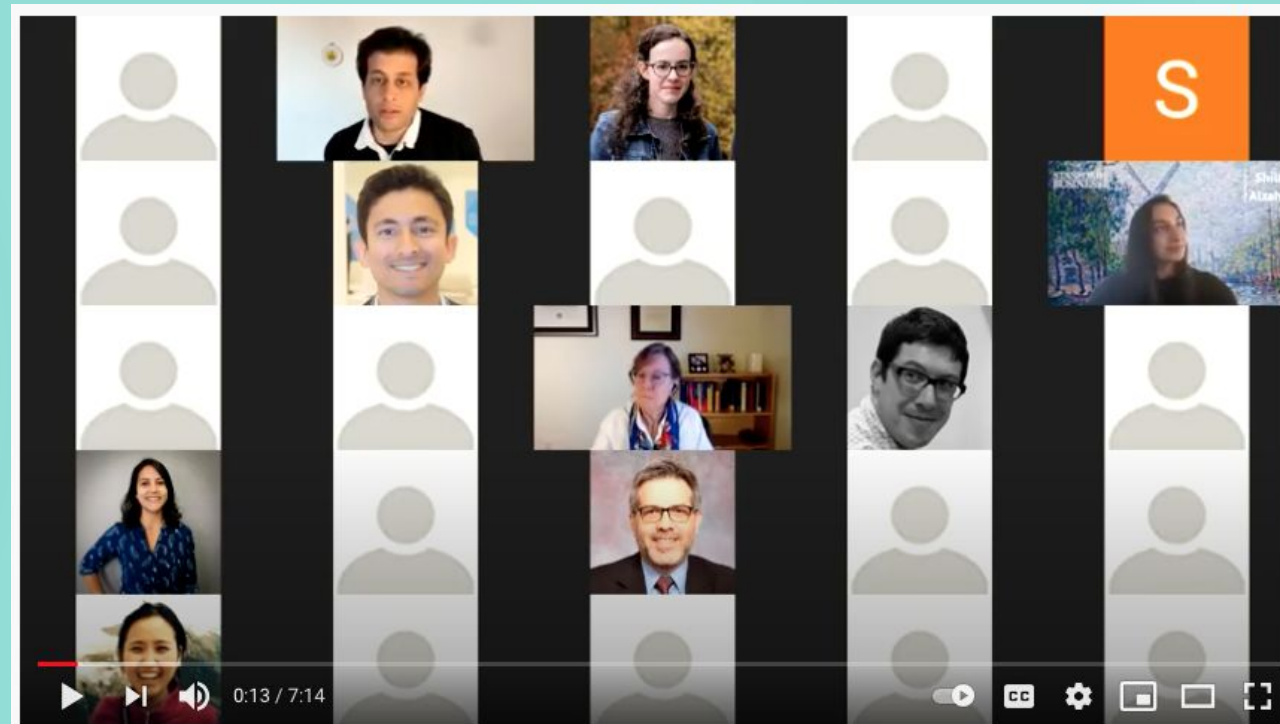
The conference has three broad areas of focus:

- **Evaluating reproducibility:** Systematically looking at the extent of reproducibility of a paper or even in a whole field is important to understand where weaknesses exist. Does, say, economics fall flat while demography shines? How should we approach these reproductions? What aspects contribute to the extent of reproducibility.
- **Practices of reproducibility:** We need new tools and approaches that encourage us to think more deeply about reproducibility and integrate it into everyday practice.
- **Teaching reproducibility:** While it is probably too late for most of us, how can we ensure that today's students don't repeat our mistakes? What are some case studies that show promise? How can we ensure this doesn't happen again?

Toronto Reproducibility Workshop

A two-day workshop focusing on reproducibility in data-centric analysis

<https://rohanalexander.com/reproducibility.html>



Toronto Reproducibility Workshop

<https://rohanalexander.com/reproducibility.html>



Sean J. Taylor @seanjtaylor · Apr 3

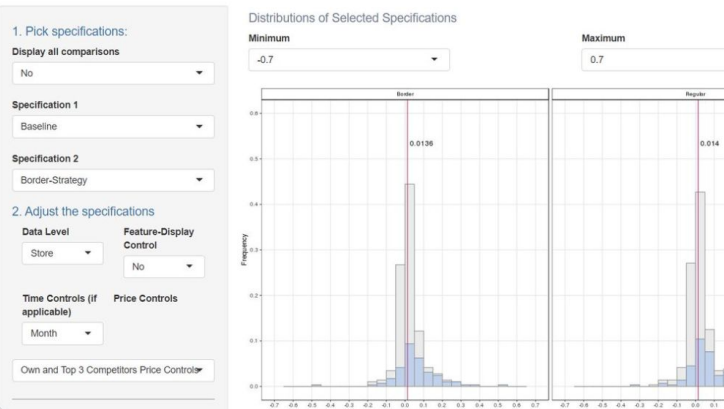
Dear Researchers, what is preventing you from presenting results like *this*?



Stephan Seiler @SeilerStephan · Apr 3

The authors even provide an interactive website that lets you explore different specifications and the robustness of their results: advertising-effects.chicagobooth.edu

[Show this thread](#)



27

30

361



Version control

with Git and GitHub

for students

- + **learn** a best practice for reproducibility
- + get familiar with systems that are widely used in industry and academia
- + facilitate collaboration and sharing

for educators

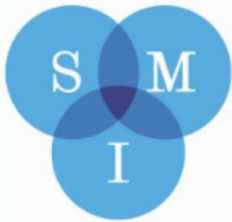
- + **teach** a best practice for reproducibility
- + centralise the distribution and collection of assignments
- + enable students to work collaboratively (even when working remotely!)

Mine Çetinkaya-Rundel

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



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

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)

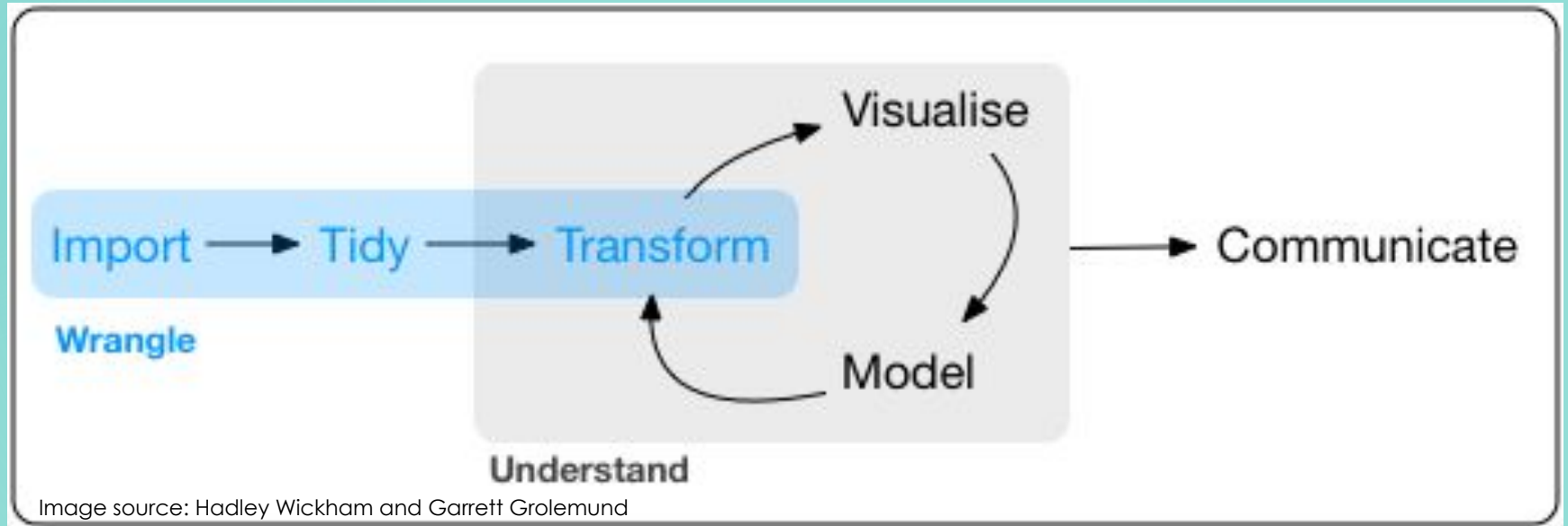
- ▶ “LUSTRE: An online data management and student project resource” (Towse et al, <https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2118645>)
- ▶ “Teaching for Large-Scale Reproducibility Verification” (Vilhuber et al, <https://www.tandfonline.com/doi/full/10.1080/26939169.2022.2074582>)
- ▶ “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>)

Back to my questions

- ▶ **Why are these methods important for students to learn?**
- ▶ What are the capacities we want them to develop?
- ▶ What are the skills that we need them to master to fluently utilize these methods?
- ▶ What are best practices for teaching these methods?
- ▶ When and where can these methods be incorporated into the K-12 and college curricula?
- ▶ How do we assess how effectively students can use these methods?

Problem-solving cycle

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



Back to my questions

- ▶ Why are these methods important for students to learn?
- ▶ **What are the capacities we want them to develop?**
- ▶ **What are the skills that we need them to master to fluently utilize these methods?**
- ▶ What are best practices for teaching these methods?
- ▶ When and where can these methods be incorporated into the K-12 and college curricula?
- ▶ How do we assess how effectively students can apply these methods?

Back to my questions

What are the capacities we want them to develop? What are the skills that we need them to master to fluently utilize these methods?

From NASEM (2018):

- ▶ workflow and reproducibility
- ▶ data management
- ▶ communication and teamwork
- ▶ ethical practice

This requires repeated exposures and reinforcement over multiple years

Back to my questions

- ▶ Why are these methods so important for all students to learn?
- ▶ What are the capacities we want them to develop?
- ▶ What are the skills that we need them to master to fluently utilize these methods?
- ▶ **What are best practices for teaching these methods?**
- ▶ When and where can these methods be incorporated into the K-12 and college curricula?
- ▶ How do we assess how effectively students can use these methods?

Back to my questions

What are best practices for teaching these methods?

- ▶ Very much a work in progress
- ▶ Development progression is important
- ▶ Need to develop, pilot, improve, then vet course materials and curricular modules
- ▶ More research is needed

Back to my questions

- ▶ Why are these methods important for students to learn?
- ▶ What are the capacities we want them to develop?
- ▶ What are the skills that we need them to master to fluently utilize these methods?
- ▶ What are best practices for teaching these methods?
- ▶ **When and where can these methods be incorporated into the K-12 and college curricula?**
- ▶ How do we assess how effectively students can use these methods?

Back to my questions

When and 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)

Back to my questions

- ▶ Why are these methods important for students to learn?
- ▶ What are the capacities we want them to develop?
- ▶ What are the skills that we need them to master to fluently utilize these methods?
- ▶ What are best practices for teaching these methods?
- ▶ When and where can these methods be incorporated into the K-12 and college curricula?
- ▶ **How do we assess how effectively students can use these methods?**

Back to my questions

How do we assess how effectively students can use these methods?

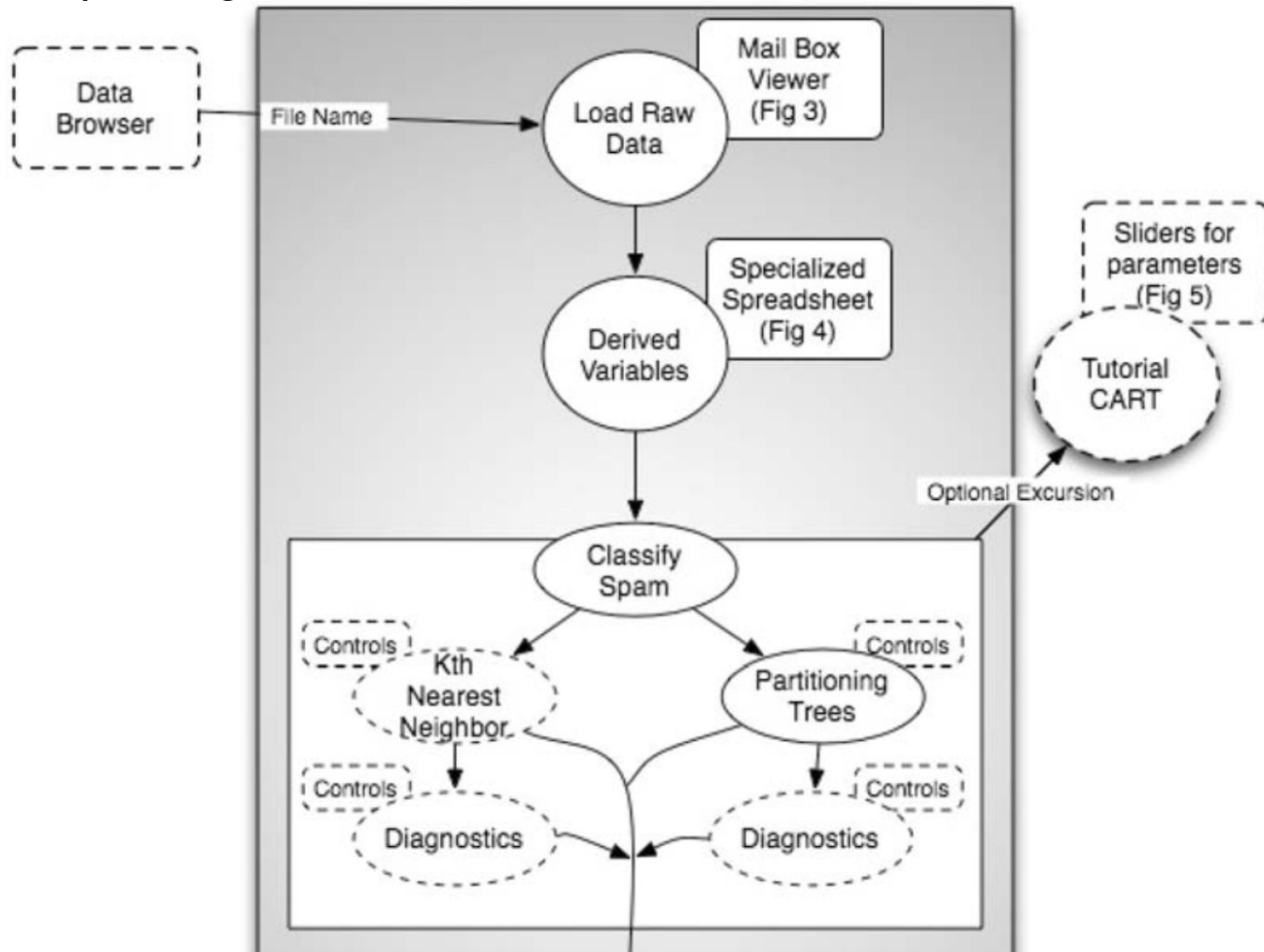
- ▶ Even more work needed here
- ▶ Start with TIER “soup to nuts” exercises
- ▶ Assessment of GitHub usage via standards based grading
- ▶ See Janz ISP (2015), <https://osf.io/hqr3j> and JSDSE special issue on teaching reproducibility

Back to the vision for dynamic documents

Dynamic, Interactive Documents for Teaching Statistical Practice

303

Nolan and Temple Lang, ISR, 2007, <https://doi.org/10.1111/j.1751-5823.2007.00025.x>



Shameless plug

- ▶ 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>

Journal of
**Statistics and
Data Science
Education**



JSDSE next steps: Data and code sharing

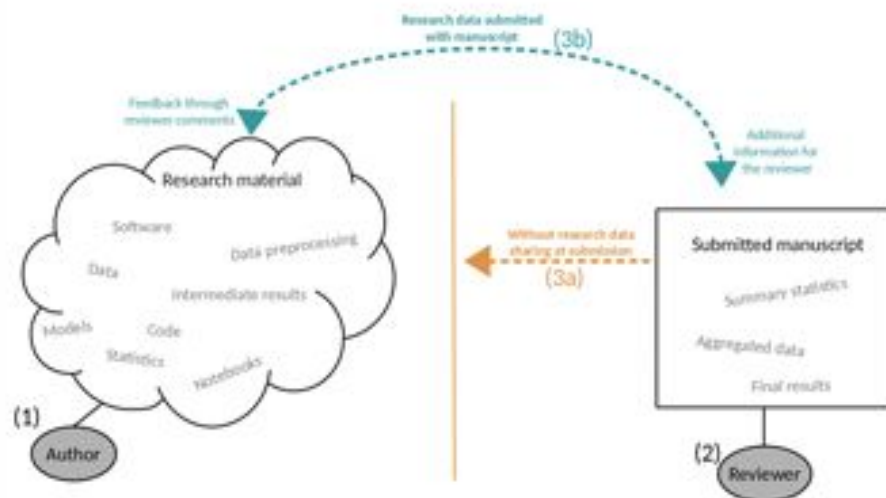
When should data and code be made available?

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

Teaching reproducibility and responsible workflows

Nicholas J. Horton, Amherst College

January 2023, nhorton@amherst.edu

```
31 def __init__(self, settings):
32     self.file = None
33     self.fingerprints = set()
34     self.logdups = True
35     self.debug = debug
36     self.logger = logging.getLogger(__name__)
37     if path:
38         self.file = open(os.path.join(path, "reports.html"),
39                         "w")
40         self.fingerprints.update(self.fingerprints)
41
42 @classmethod
43 def from_settings(cls, settings):
44     debug = settings.getbool("debug")
45     return cls(job_dir(settings), debug)
46
47 def request_seen(self, request):
48     fp = self.request_fingerprint(request)
49     if fp in self.fingerprints:
50         return True
51     self.fingerprints.add(fp)
52     if self.file:
53         self.file.write(fp + os.linesep)
54
55 def request_fingerprint(self, request):
56     return request_fingerprint(request)
```

Image source: Wikicommons



Image source: heylagostechie

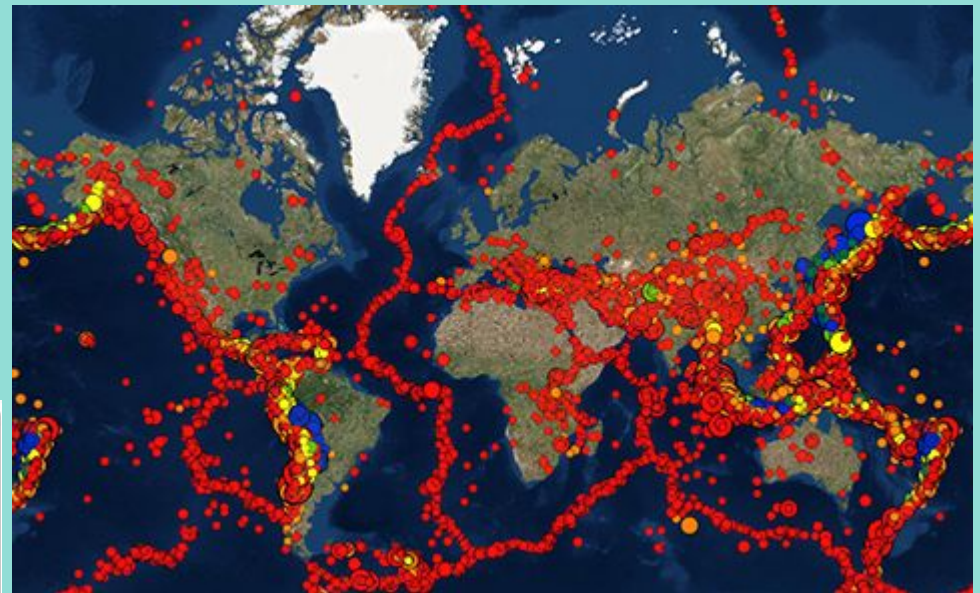


Image source: Concord Consortium

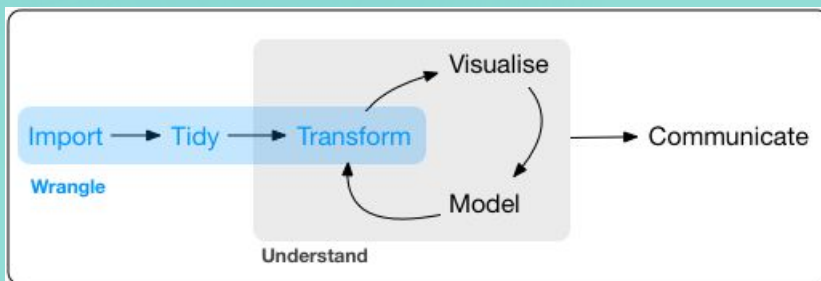


Image source: Hadley Wickham and Garrett Grolmund

TIER symposium: Nicole Janz

- ▶ “Teaching replication” keynote
- ▶ Qualitative replication as a pedagogical approach
<https://doi.org/10.1017/S1049096520000864>

Guidance From Course Instructors

Clear aim:

Are students conducting a replication or duplication?

Be transparent & reproducible:

1. **Selection:** How will they select the original study for replication?
2. **Pre-register:** Can they avoid accusations of error hunting?
3. **Cross-check:** Who will cross-check the replication results before reporting them?
4. **Authors:** Will they contact the original authors and can you help them with an email template?
5. **Publication:** Do students plan journal submission or is this for learning purpose only?

@polscreplicate



TIER symposium: Amelia McNamara

- ▶ “Consistency is key: a case study in R syntax”
- ▶ Take home messages:
 - ▶ using tools like RMarkdown and RStudio instructors can scaffold more for their students (and they can get further faster)
 - ▶ code reading can foster deeper understanding

Tidyverse syntax

```
data %>% goal(x)
```

SUMMARY STATISTICS:

one continuous variable:

```
mtcars %>% dplyr::summarize(mean(mpg))
```

one categorical variable:

```
mtcars %>% dplyr::group_by(cyl) %>%  
dplyr::summarize(n())
```

the pipe

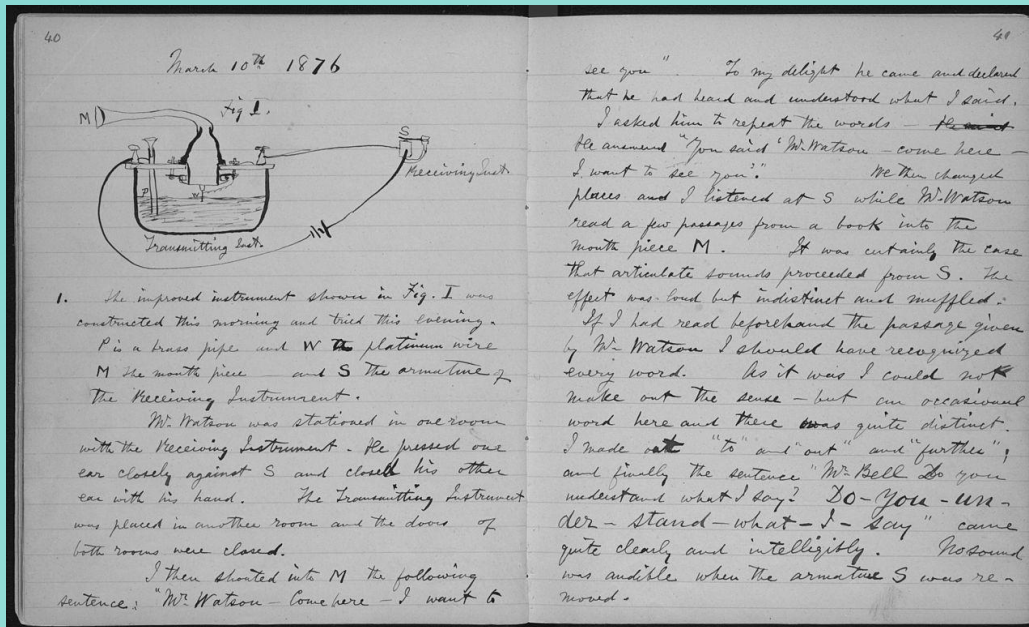
two categorical variables:

```
mtcars %>% dplyr::group_by(cyl, am) %>%  
dplyr::summarize(n())
```




TIER symposium: Nicholas Bussberg

- ▶ “Incorporating an accessible reproducibility workflow into entry-level courses”
- ▶ lab notebook metaphor
- ▶ reproducibility in all courses: start simple and build on a common framework so that it is expected



TIER symposium: Jessica Sullivan

- ▶ “A practical approach to teaching reproducibility, and improving your own research”
- ▶ “do-it-again-ability”
- ▶ find some in-press or recent papers and re-enact the study (without the kids)

 Jessica Sullivan: "A practical approach to teaching reproducibility & improv..."

Write a method section

1. Classic version: write and peer-edit
2. Modified version:
 - a. Edit my in-prep work
 - b. Watch participant videos and write methods
 - c. Write methods based on the datasheet used to collect data
 - d. Read reviews and response to reviews from my under review work



TIER symposium: Phil McAleer

- ▶ “Creating a curriculum centered on reproducible research”
- ▶ previous insight: students were fine if the data were in the right format. But not otherwise
- ▶ Introduce, reinforce, work towards mastery

<https://psyteachr.github.io/>



Our curriculum now emphasizes **essential ‘data science’ graduate skills** that have been overlooked in traditional approaches to teaching, including programming skills, **data visualisation**, **data wrangling** and **reproducible reports**. Students learn about probability and inference through data simulation as well as by working with **real datasets**.

#PSYTEACHR REPRODUCIBLE RESEARCH



Working with data and research in a reproducible fashion



TIER symposium: Rachel Hayes-Harb

Reproducibility Education in an Undergraduate Capstone Course

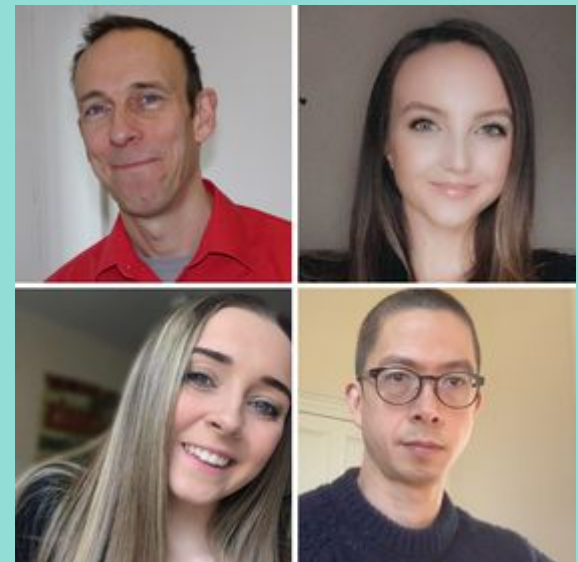
- ▶ I will discuss the development, implementation, and learning outcomes assessment associated with an undergraduate capstone course I teach
- ▶ The primary goals of the course are to provide a meaningful and authentic research experience for undergraduate students, and to do so by embedding research skills development within responsible conduct of research and Open Science values and practices.



TIER symposium: Towse, Davies, James, and Ball

LUSTRE: An online tool for training students in data management and data sharing

- ▶ LUSTRE, the Lancaster University STatistics REsource is a student project data management system is designed to deploy open science practices amongst psychology students working with empirical data.
- ▶ LUSTRE is an open source, online data catalogue system, that captures key data management information about a student research project.



TIER symposium: Sam Parsons and Flavio Azevedo

- ▶ **Building a community from open scholarship pedagogy with a Framework for Open and Reproducible Research Training (FORRT)**
 - ▶ FORRT Educational Nexus, which combines eight distinct initiatives aimed at promoting the integration of open and reproducible science into higher education.
 - ▶ FORRT pedagogies, which shares exemplary instances of teaching open and reproducible research practices, including the valuable pedagogies behind the teaching materials.

FORRT is an organised community effort to curate and evaluate educational outcomes of open scholarship reforms, as a pedagogy based route to improve research practices.



TIER symposium: Fernando de la Guardia

How to Teach Reproducibility in the Classroom (BITSS)

The Berkeley Initiative for Transparency in the Social Sciences (BITSS) has developed an adaptable curricular module to teach reproducible research through reproductions of published work.



9 Code Review Best Practices



Source:kinsta.com

1. Know what to look for in a code review
2. Build and test (before review)
3. Don't review for longer than 15 minutes
4. Check no more than 400 lines at a time (smaller PR are better?)
5. Give feedback that helps (not hurts)
6. Communicate goals and expectations
7. Include everyone in the code review process
8. Foster a positive culture
9. Automate to save time

<https://www.perforce.com/blog/qac/9-best-practices-for-code-review>

Next Generation Science Standards (NGSS, 2013)

See for example: MS-LS2-1 Ecosystems: Interactions, Energy, and Dynamics

Students who demonstrate understanding can:

- MS-LS2-1.** **Analyze and interpret data to provide evidence for the effects of resource availability on organisms and populations of organisms in an ecosystem.** [Clarification Statement: Emphasis is on cause and effect relationships between resources and growth of individual organisms and the numbers of organisms in ecosystems during periods of abundant and scarce resources.]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices

Analyzing and Interpreting Data

Analyzing data in 6–8 builds on K–5 experiences and progresses to extending quantitative analysis to investigations, distinguishing between correlation and causation, and basic statistical techniques of data and error analysis.

- Analyze and interpret data to provide evidence for phenomena.

Disciplinary Core Ideas

LS2.A: Interdependent Relationships in Ecosystems

- Organisms, and populations of organisms, are dependent on their environmental interactions both with other living things and with nonliving factors.
- In any ecosystem, organisms and populations with similar requirements for food, water, oxygen, or other resources may compete with each other for limited resources, access to which consequently constrains their growth and reproduction.
- Growth of organisms and population increases are limited by access to resources.

Crosscutting Concepts

Cause and Effect

- Cause and effect relationships may be used to predict phenomena in natural or designed systems.

Connections to other DCIs in this grade-band:

MS.ESS3.A ; MS.ESS3.C

Articulation of DCIs across grade-bands:

3.LS2.C ; 3.LS4.D ; 5.LS2.A ; HS.LS2.A ; HS.LS4.C ; HS.LS4.D ; HS.ESS3.A

Common Core State Standards Connections:

ELA/Literacy -

RST.6-8.1 Cite specific textual evidence to support analysis of science and technical texts. (MS-LS2-1)

RST.6-8.7 Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). (MS-LS2-1)

<https://www.nextgenscience.org>

Next Generation Science Standards (NGSS, 2013)

Science and Engineering Practices

Analyzing and Interpreting Data

Analyzing data in 6–8 builds on K–5 experiences and progresses to extending quantitative analysis to investigations, distinguishing between correlation and causation, and basic statistical techniques of data and error analysis.

- Analyze and interpret data to provide evidence for phenomena.

Common Core State (Math) Standards Connections:

RST.6-8.1 Cite specific textual evidence to support analysis of science and technical texts. (MS-LS2-1)

RST.6-8.7 Integrate quantitative or technical information expressed in words in a text with a version of that information expressed visually (e.g., in a flowchart, diagram, model, graph, or table). (MS-LS2-1)