Data Management:
The First Step in Reproducible Research

Harvard Chan Bioinformatics Core | Tools for Reproducible Research

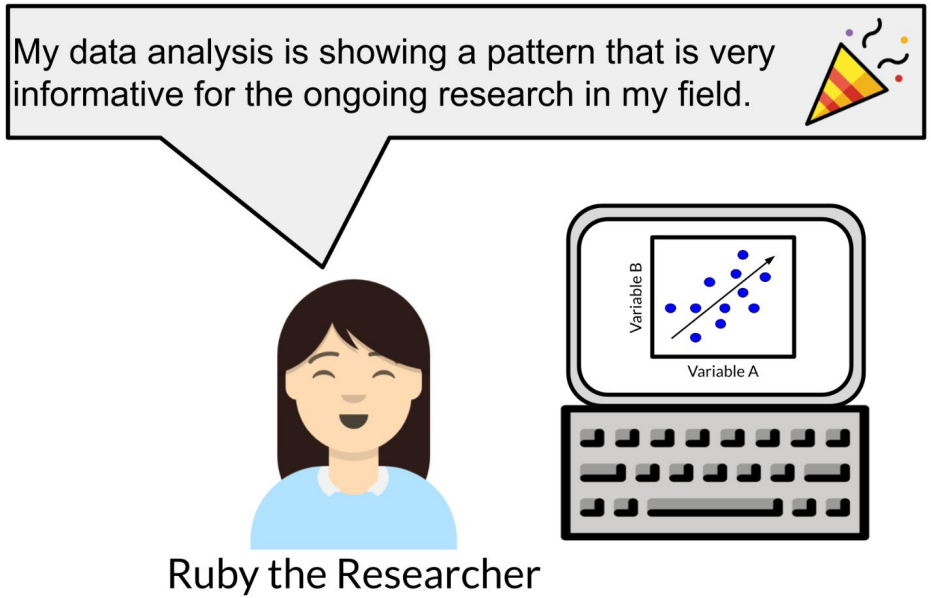
October 30, 2026

Julie Goldman | julie_goldman@harvard.edu

Learning Objectives

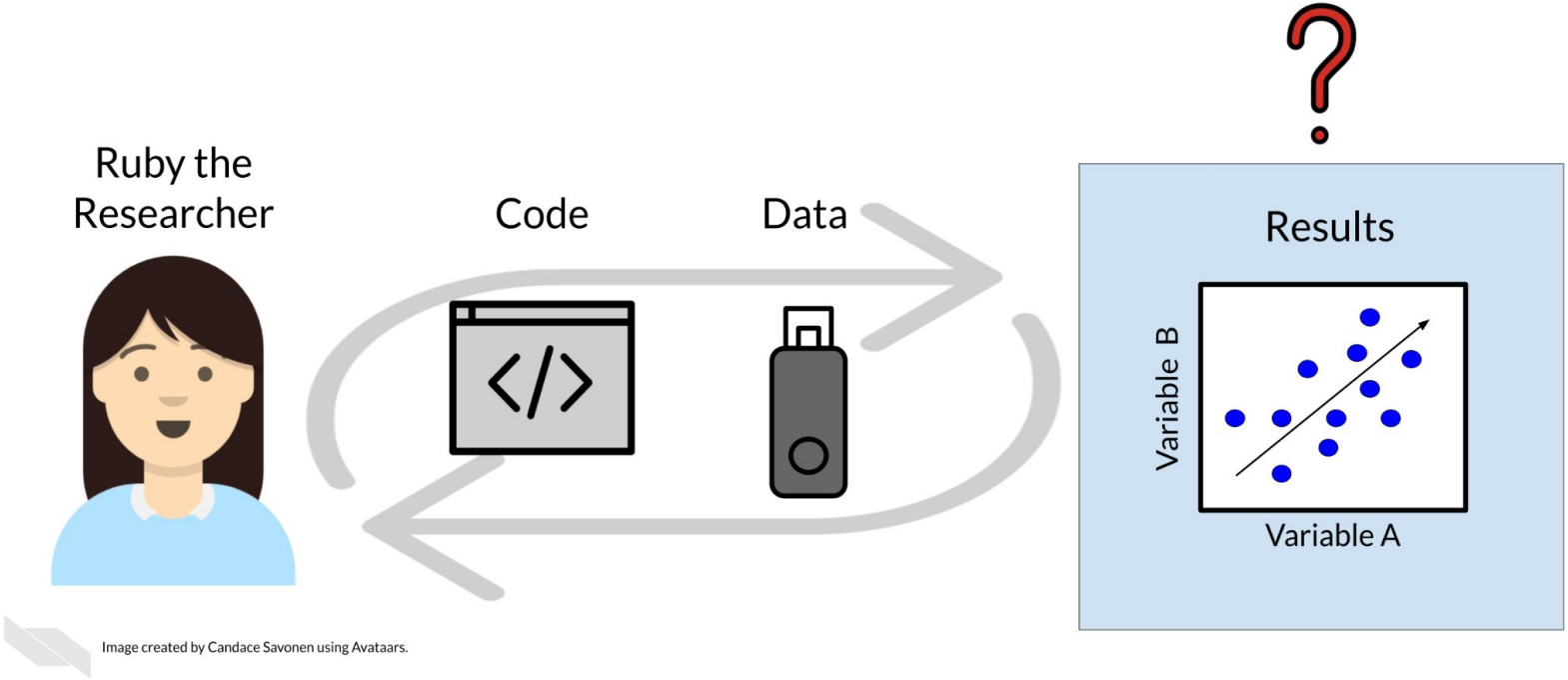
- Understand the impact of creating reproducible research
- Examine challenges of creating reproducible research data
- Discuss foundational data management practices
- Review available tools that facilitate reproducible research data

Defining Reproducibility



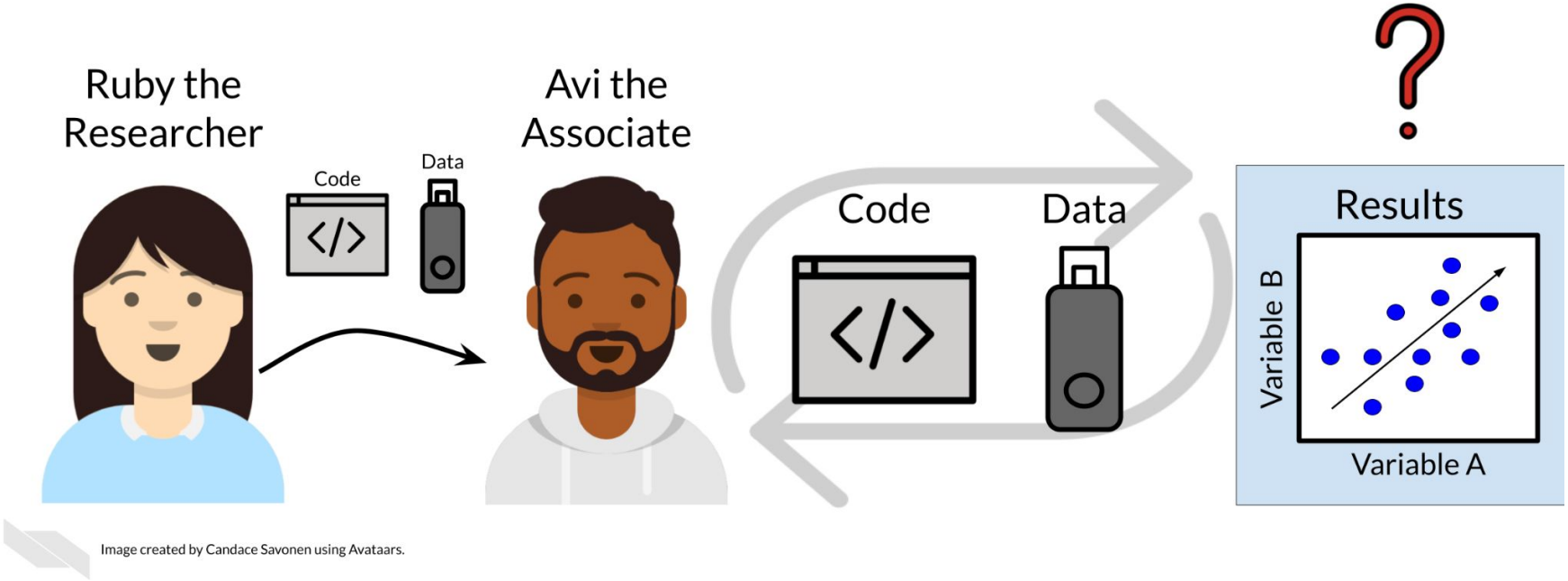
Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Defining reproducibility. Informatics Training Network

Repeatability



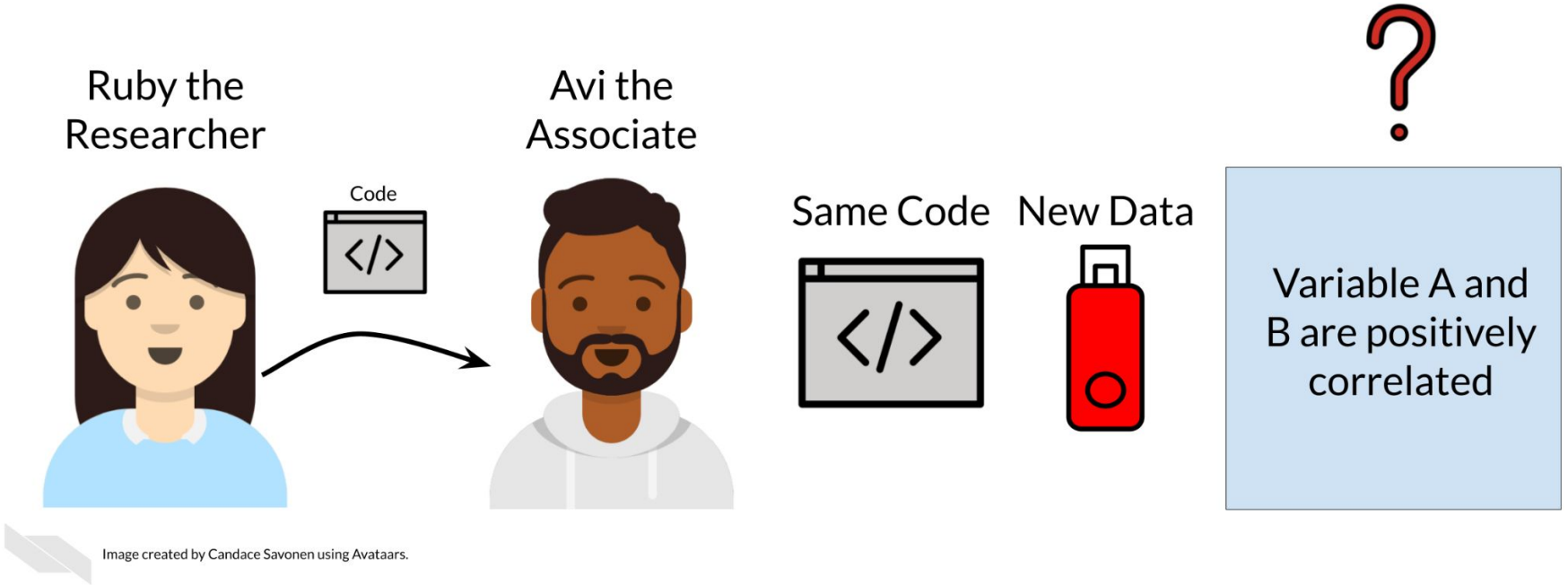
Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Defining reproducibility. Informatics Training Network

Reproducibility



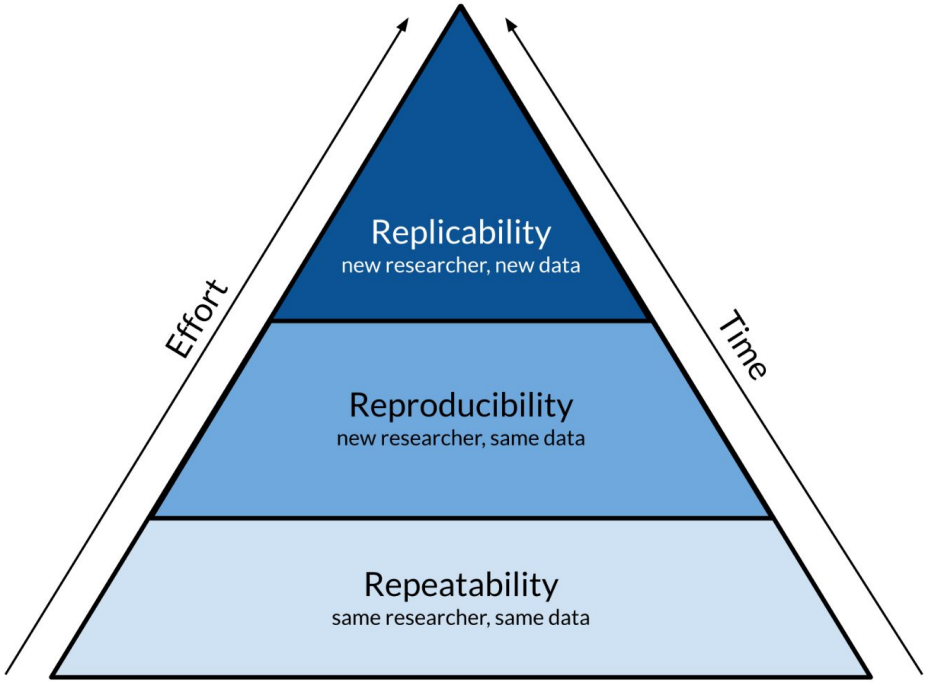
Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Defining reproducibility](). Informatics Training Network

Replicability



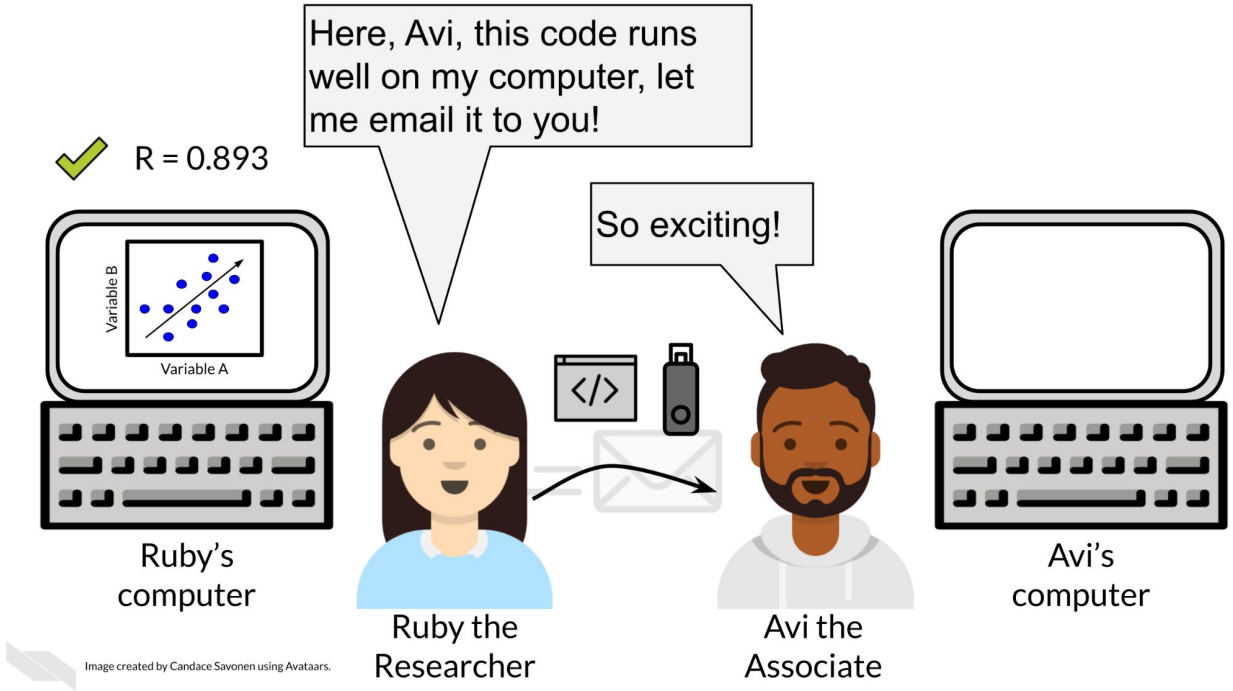
Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Defining reproducibility](). Informatics Training Network

Research process as a hierarchy



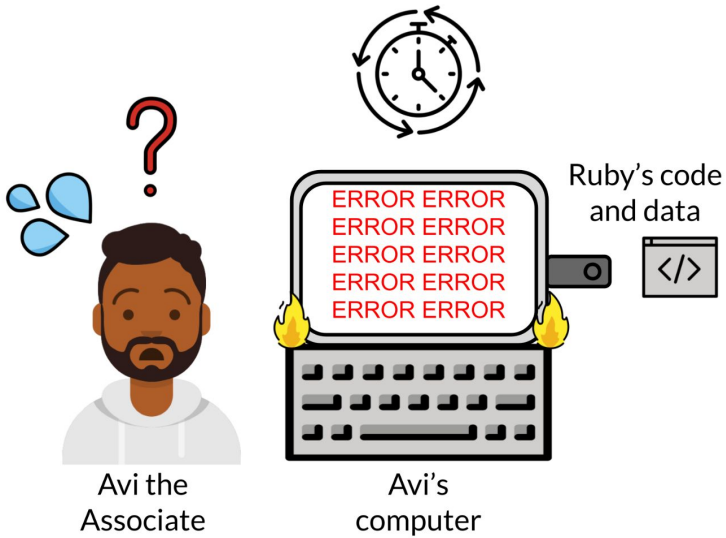
Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Defining reproducibility. Informatics Training Network

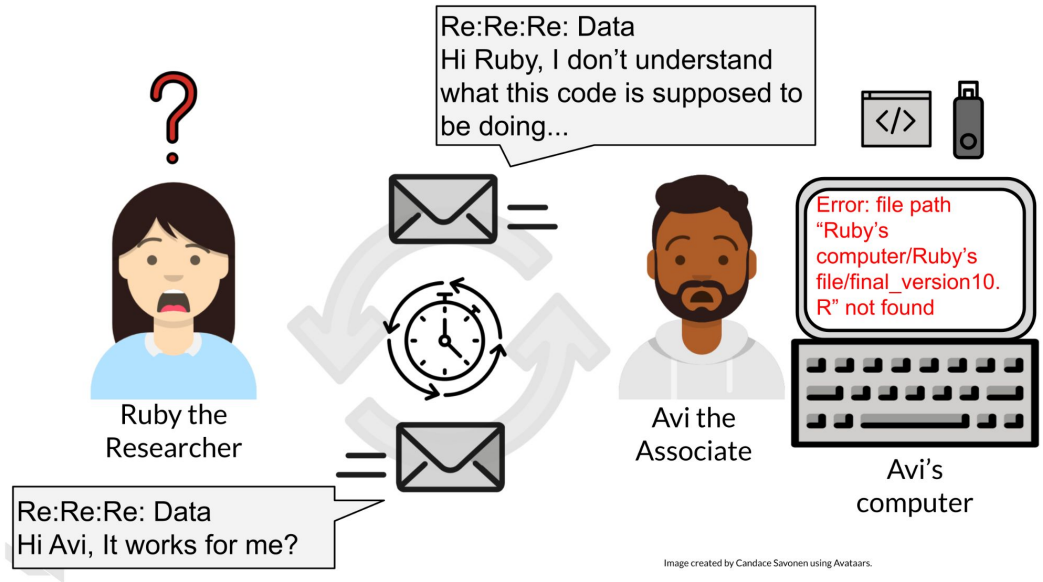
So, what's the issue?



Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Defining reproducibility](). Informatics Training Network

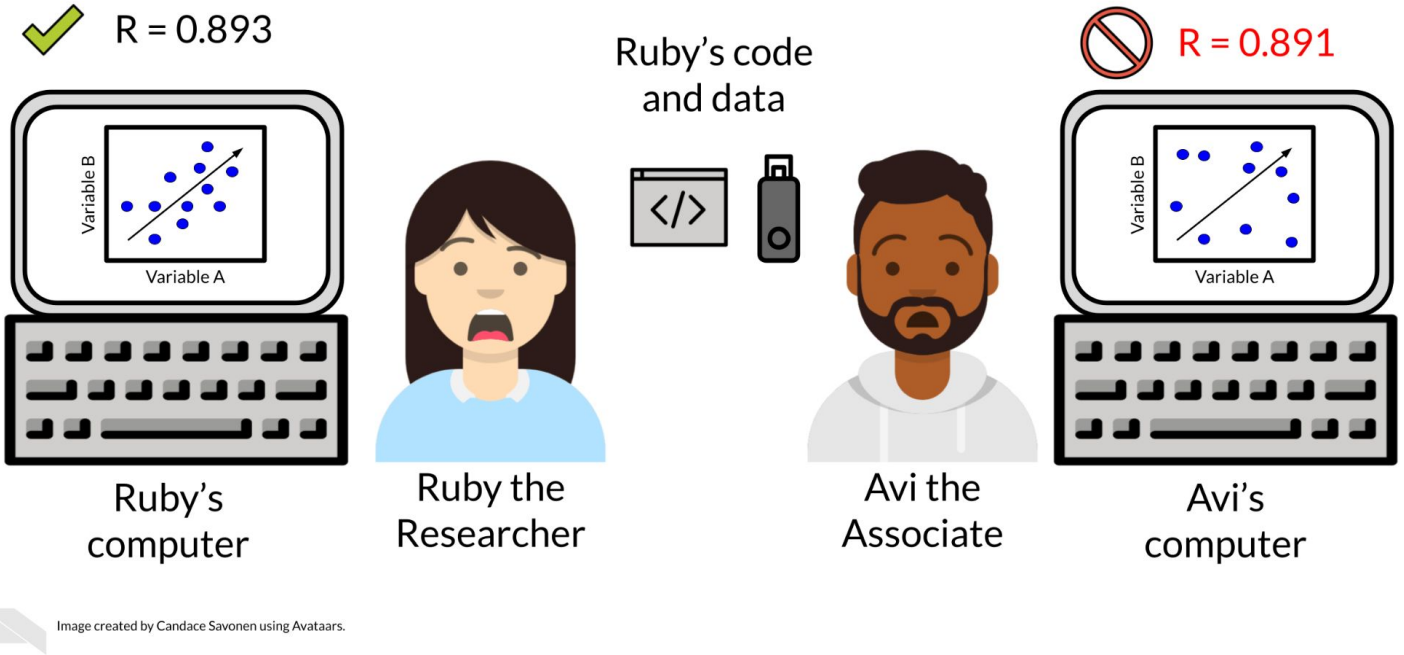
Reproducibility in daily life





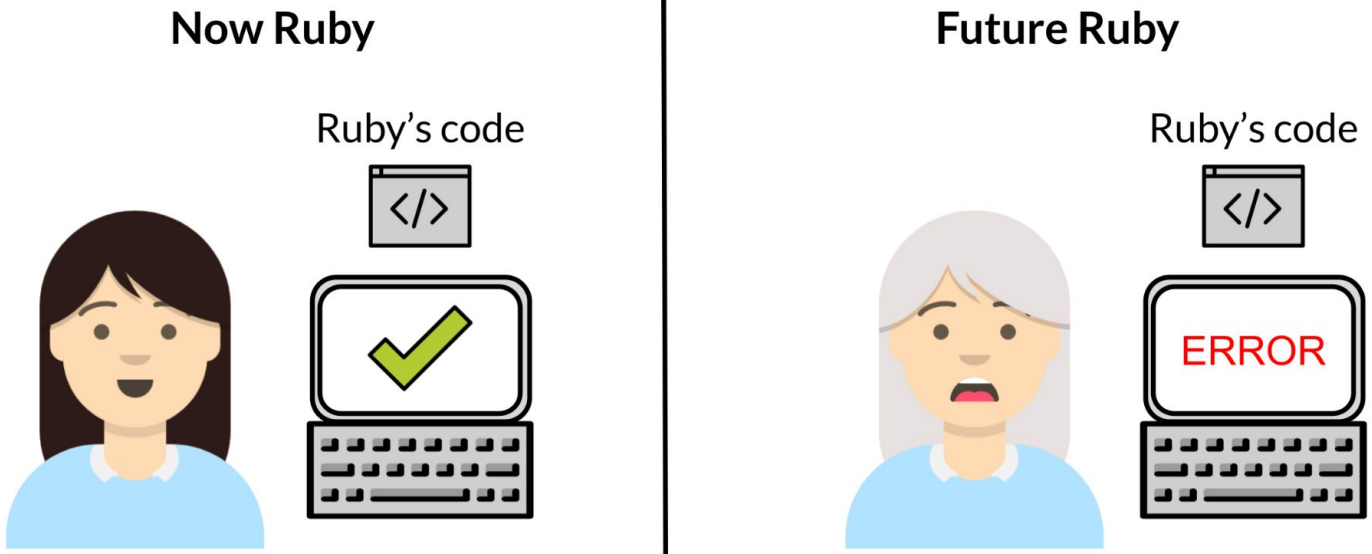
Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Defining reproducibility](). Informatics Training Network

Reproducibility in daily life



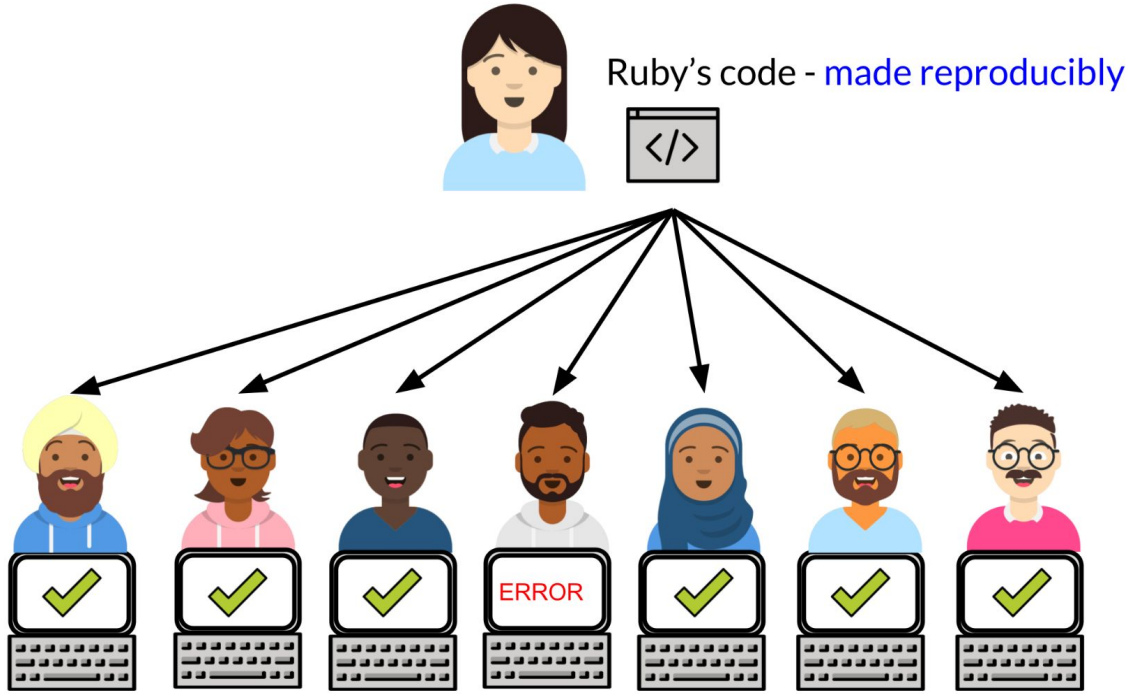
Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Defining reproducibility](). Informatics Training Network

Reproducibility in daily life



Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Defining reproducibility. Informatics Training Network

Reproducibility is worth the effort!



Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Defining reproducibility. Informatics Training Network

So, why not put in the effort?



I don’t have enough time



Technical obsolescence



I don’t have the skills



There’s not enough incentive



I don’t know where to start



Another reason

Emoji Icons: All icons by OpenMoji used under the CC BY-SA 4.0 license.

You can't have any sort of reproducibility without good data and project management.

Research Data Management

Is the active and ongoing management of data through its lifecycle of interest and usefulness.

Ensures and facilitates the timely collection of complete and accurate protocol-required information.

Includes decisions that are agreed to at the beginning of a study and carried out to completion.



Research Data Lifecycle by LMA RDMWG

Data Through the Research Lifecycle

Data examples created throughout the research data lifecycle

Are final data available for validation, reproduction or reuse?

Final Data

Summary dataset, figures, publication-ready images



Raw Data

Instrument output; raw microscopy images

What about experimental methods and measurement parameters?

Intermediate Data

Normalized measurements; processed images; calculated values

Research Data Lifecycle by LMA RDMWG: CC BY-NC
Emoji Icons: All icons by OpenMoji used under the CC BY-SA 4.0 license.

Data Management Practices for Reproducibility



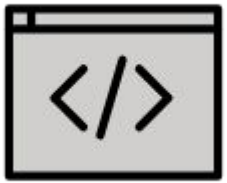
Organization

- Directory structure
- File naming
- Version control



Documentation

- README File
- Data Dictionary
- Metadata



Automation

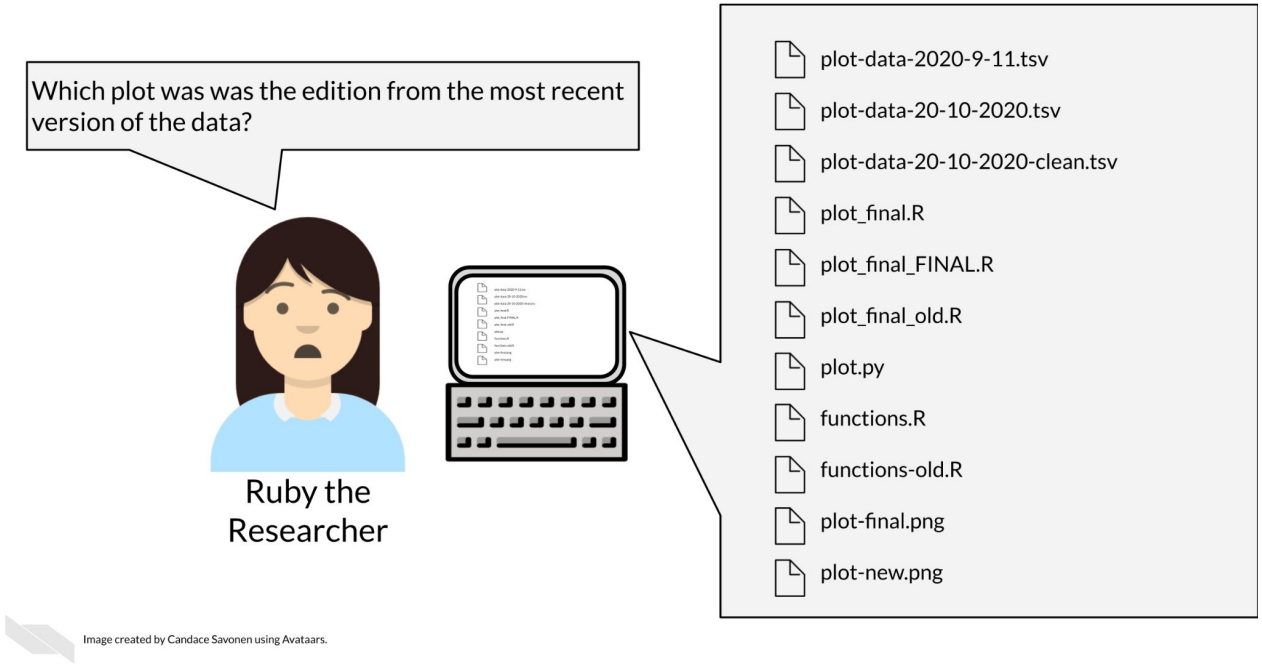
- Scripts for workflows
- Computing environment
- Dependencies



Dissemination

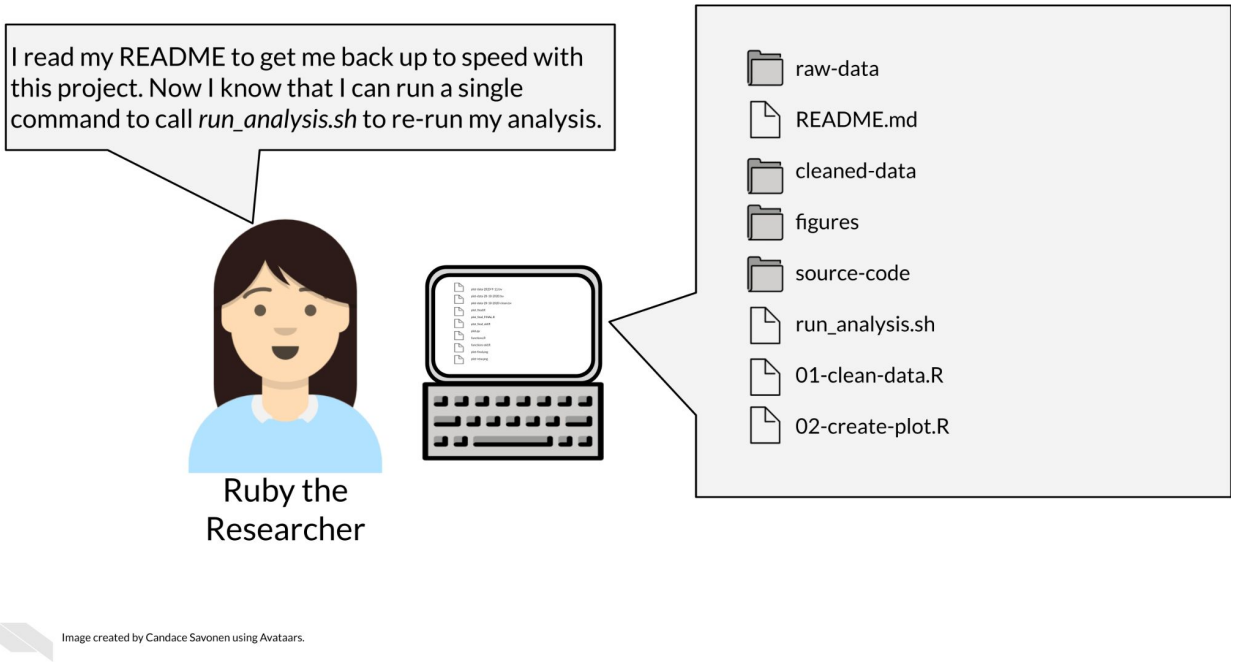
- Share in repository
- Get DOI for citation
- License and terms of use language

Organization: What to avoid



Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Organizing your project](). Informatics Training Network

Organization: Better practice

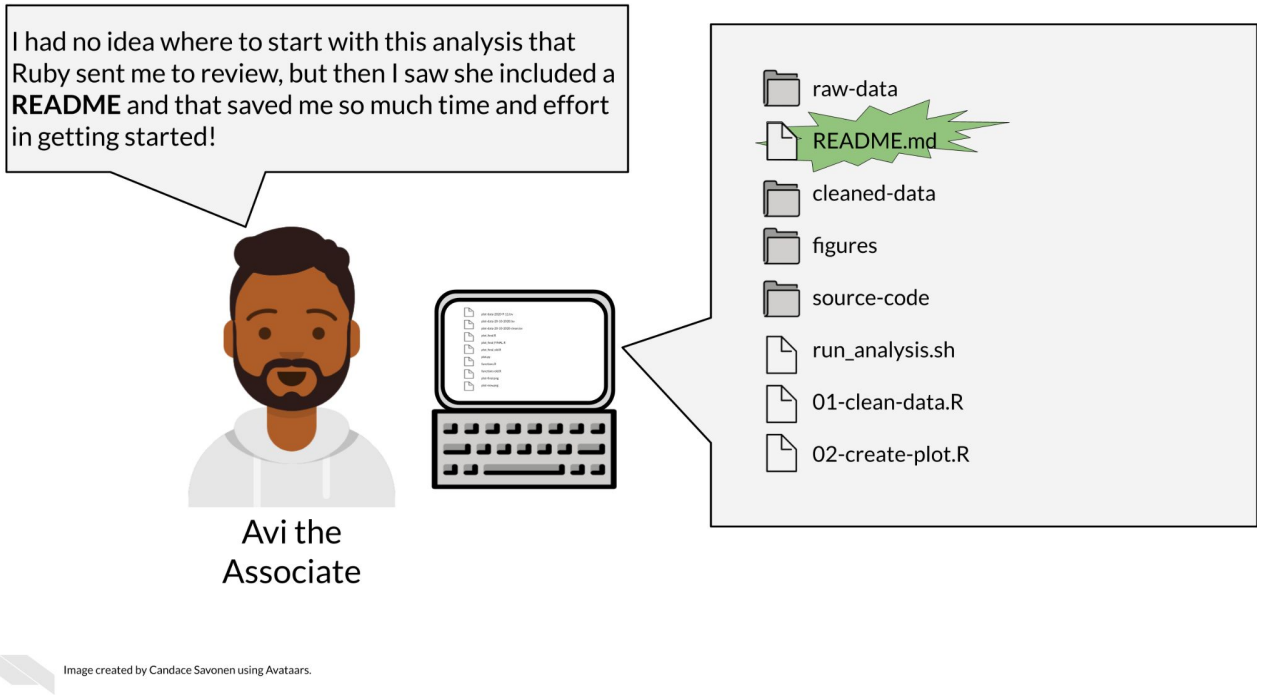


Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Organizing your project](). Informatics Training Network

Organization: Tips and tricks

- Make file names informative – avoid using spaces, quotes, or unusual characters
- Keep like-files together in their own directory – keep raw data separate from processed data or other results!
- Number scripts in the order that they are run
- Put source scripts and functions in their own directory
- Put output in its own directories like results and plots
- Have a central document (like a README) that describes the basic information about the project and analysis (see: documentation)
- Make a central script that re-runs everything (see: automation)

Documentation: Good practice



Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Documenting analyses](). Informatics Training Network

Documentation: Good practice

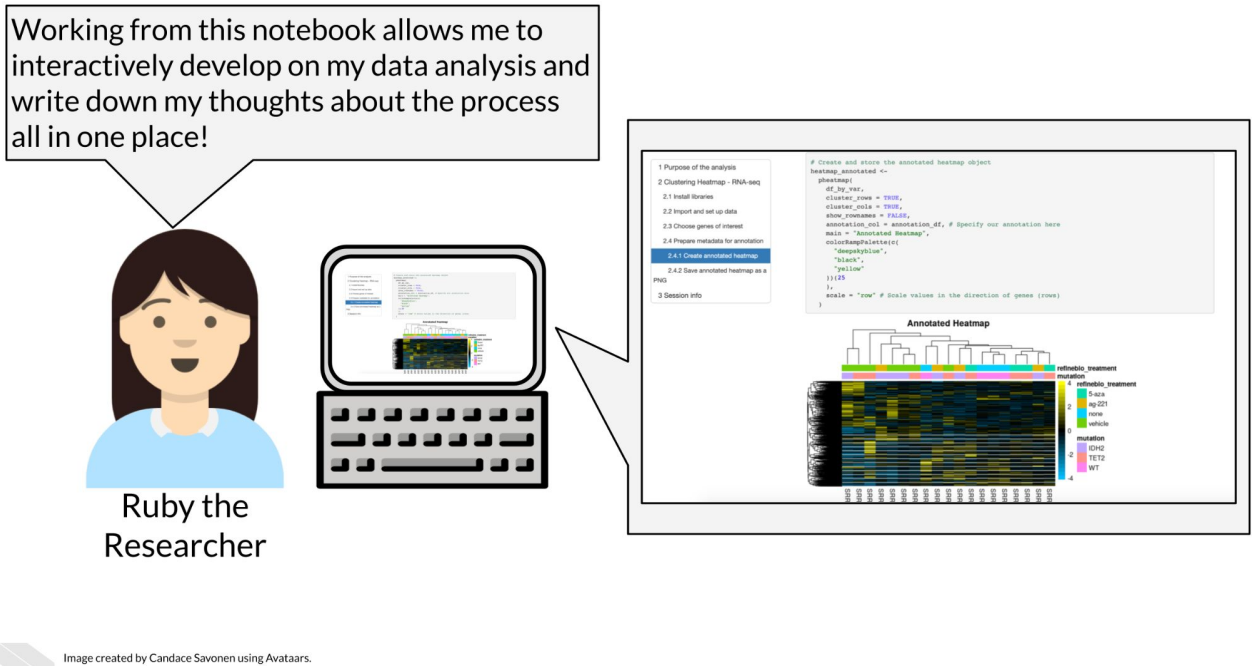
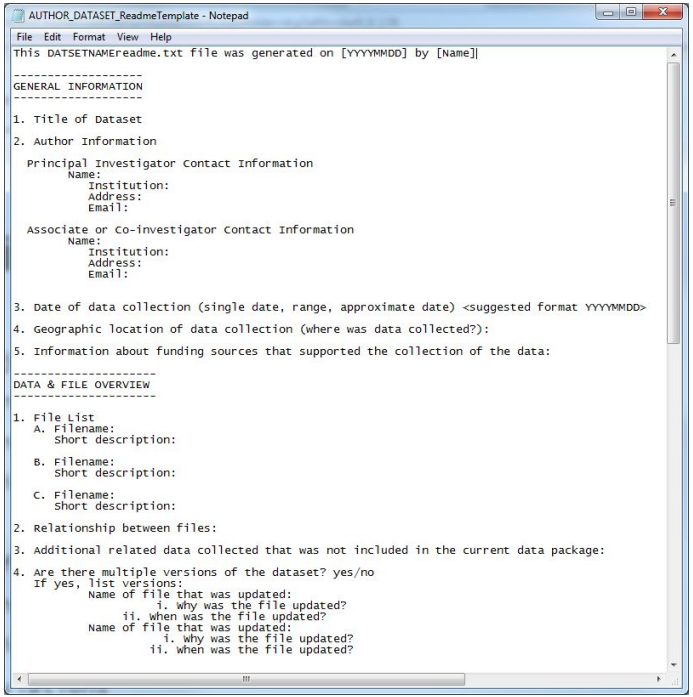
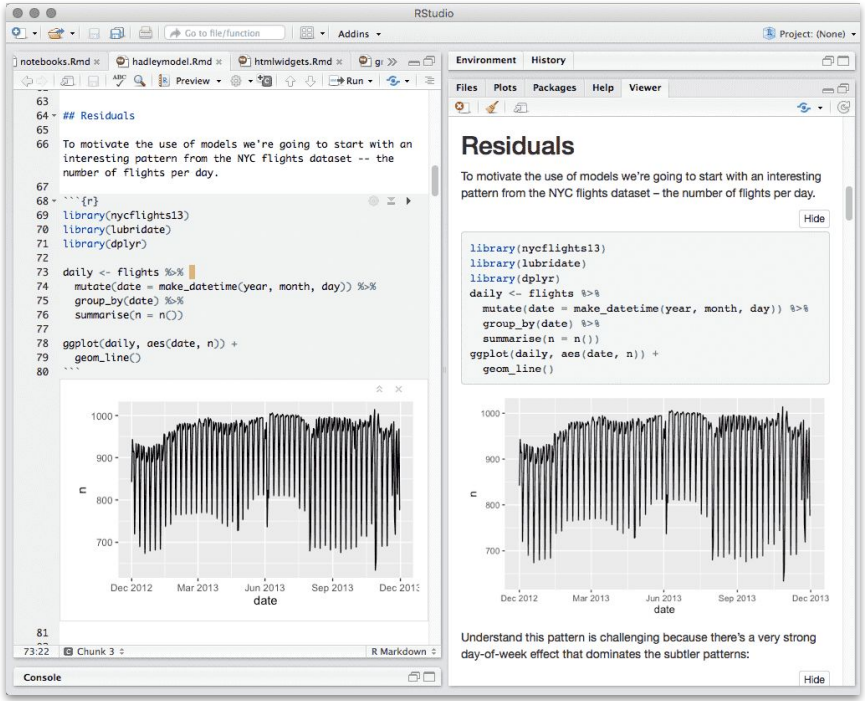


Image created by Candace Savonen using Avataars.

Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Documenting analyses](). Informatics Training Network

Documentation: Useful tools

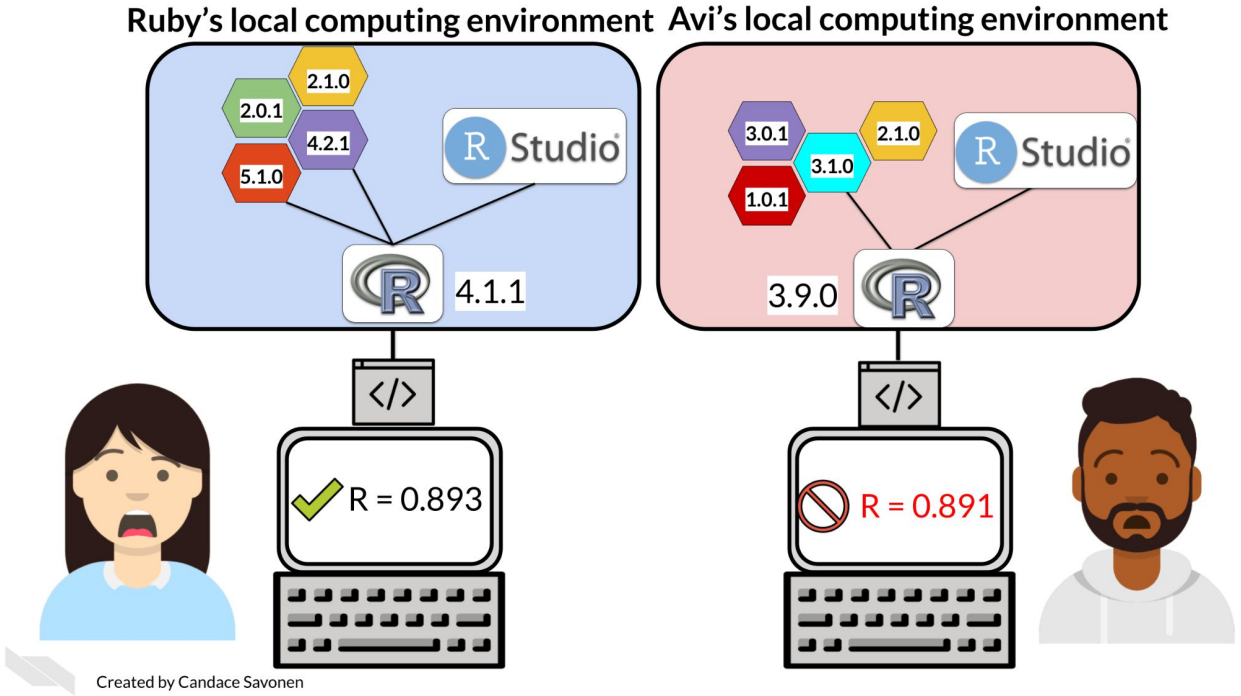




README File Example Template:
http://data.research.cornell.edu/content/readme

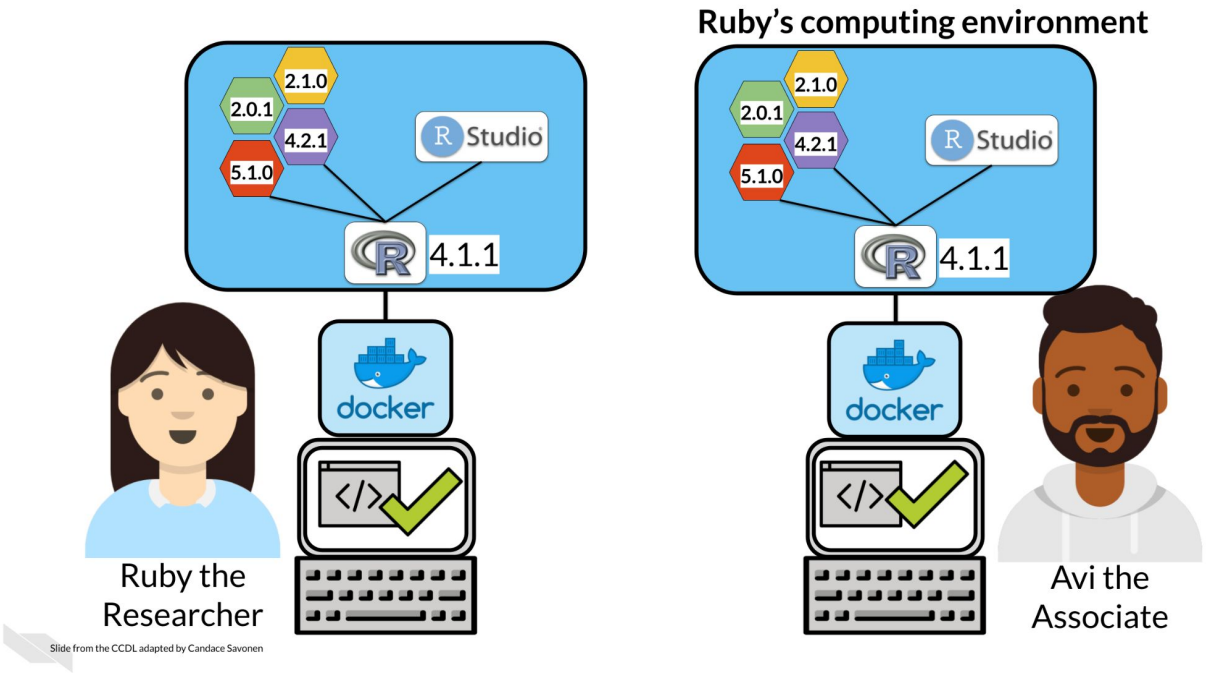
R Markdown: The Definitive Guide:
https://bookdown.org/yihui/rmarkdown

Automation: What to avoid



Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Managing package versions](). Informatics Training Network

Automation: Good practice



Source: Candace Savonen, et al. (2021). [Intro to Reproducibility in Cancer Informatics: Managing package versions](). Informatics Training Network

Automation: Tips and tricks

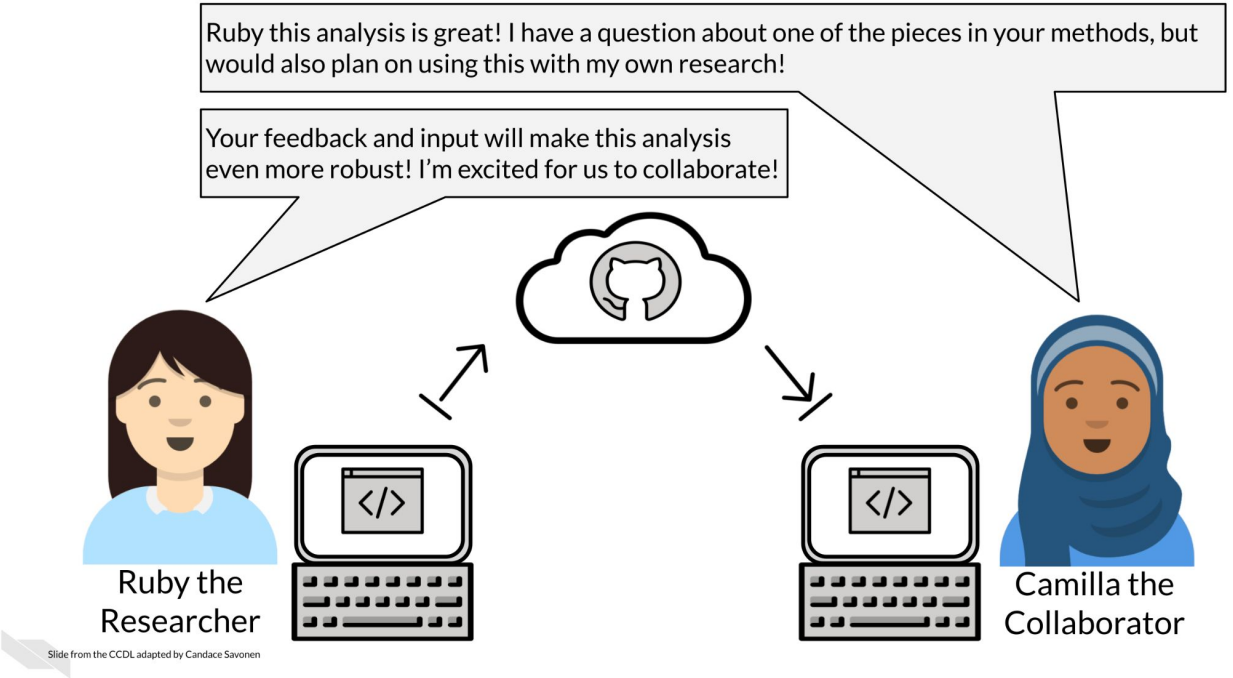
Create a script that can execute all of the various subcomponents of the entire workflow.

This simple example has three steps that can be performed automatically:

1. **`clean_data.R`** to generate the cleaned data table
2. **`analysis.R`** to perform the statistical test
3. **`runall.sh`** saved in the src directory to run the entire workflow process

```
|-- tomato_project
|   |-- data_raw
|   |   |-- raw_yield_data.csv
|   |   |-- README.txt
|   |-- src
|   |   |-- analysis.R
|   |   |-- clean_data.R
|   |   |-- runall.sh
```

Dissemination: Good practice



Source: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Making your project open source with GitHub. Informatics Training Network

Dissemination: Better practice

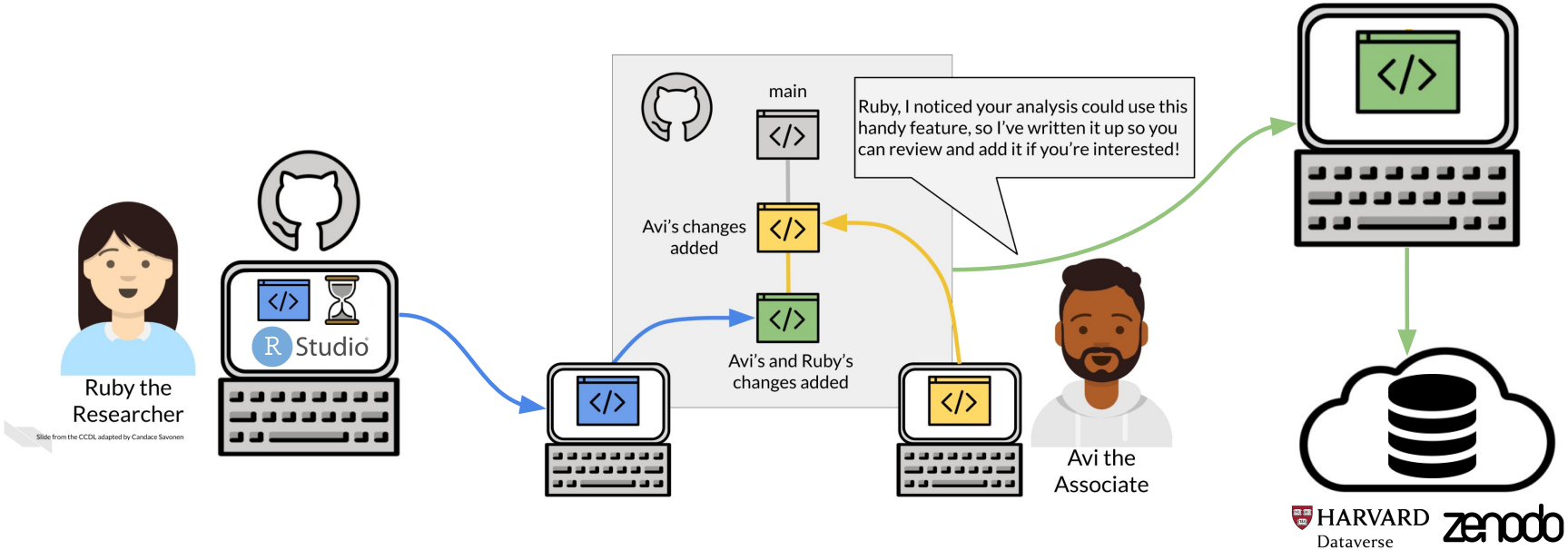
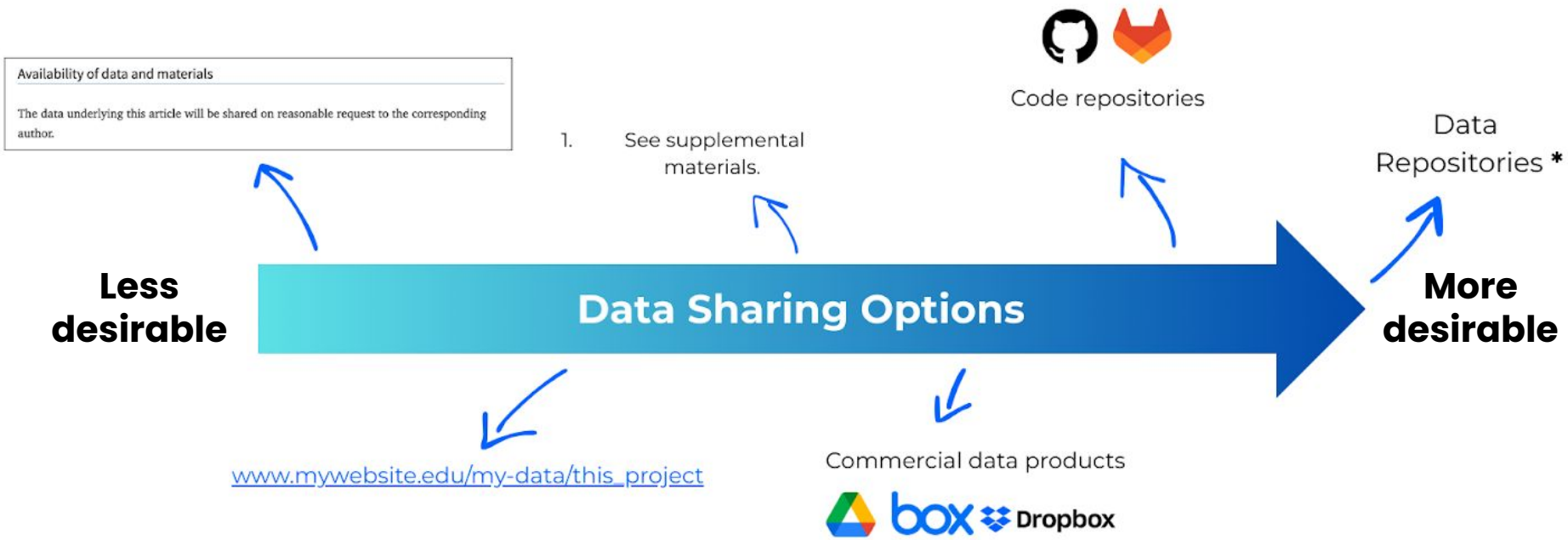


Illustration based on: Candace Savonen, et al. (2021). Intro to Reproducibility in Cancer Informatics: Making your project open source with GitHub. Informatics Training Network

Rethinking Data Sharing Practices

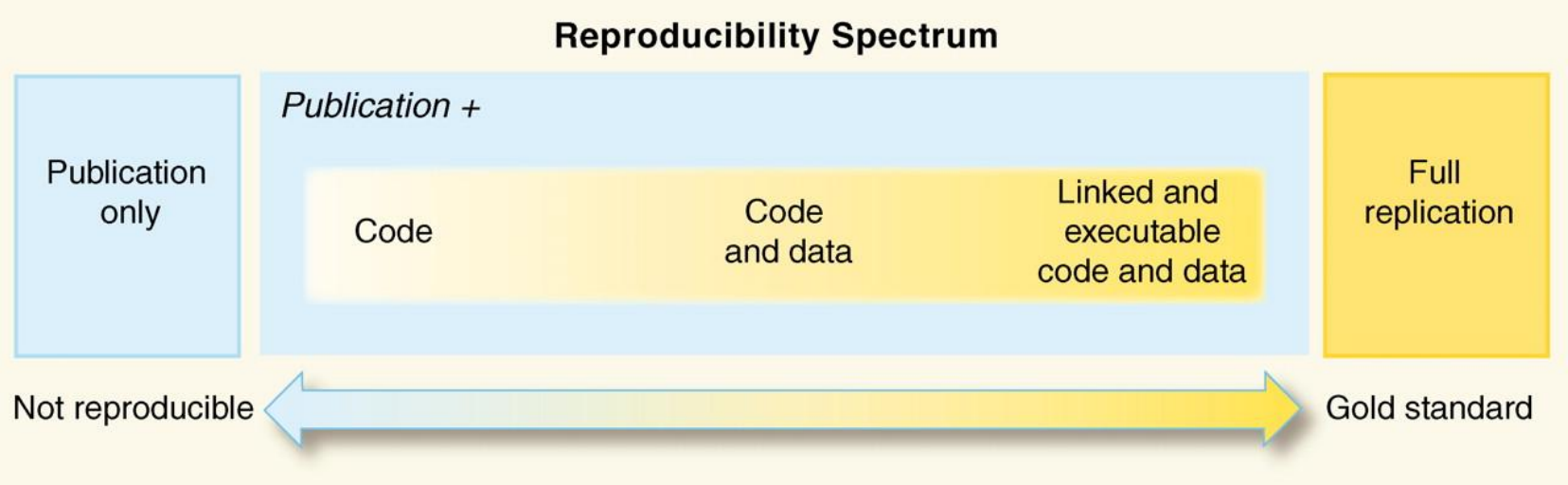


* Keep in mind the differences between disciplinary, federal/public, generalist, and institutional repositories

Dissemination: Useful tools

Disciplinary	General	Software	Methods
dbGaP GENOTYPES and PHENOTYPES	HARVARD Dataverse	GitHub	protocols.io
GEO Gene Expression Omnibus	zenodo	GitLab	nature protocols
NIMH National Institute of Mental Health	figshare	Bitbucket	bio-protocol Exchange

Putting it all together!



Source: IPeng, Roger D. 2011. "Reproducible Research in Computational Science." Science 334 (6060): 1226-1227. https://doi.org/10.1126/science.1213847

It takes some effort to organize your research to be reproducible...the principal beneficiary is generally the author themself.

– Jon Claerbout

Making Scientific Contributions Reproducible

Why Reproducibility? Think Selfishly!

Article | Open access | Published: 26 June 2024

Metabotropic signaling within somatostatin interneurons controls transient thalamocortical inputs during development

Deepanjali Dwivedi, Dimit
Qing Xu, Samuel A. Liebr
Pouchelon

Nature Communications

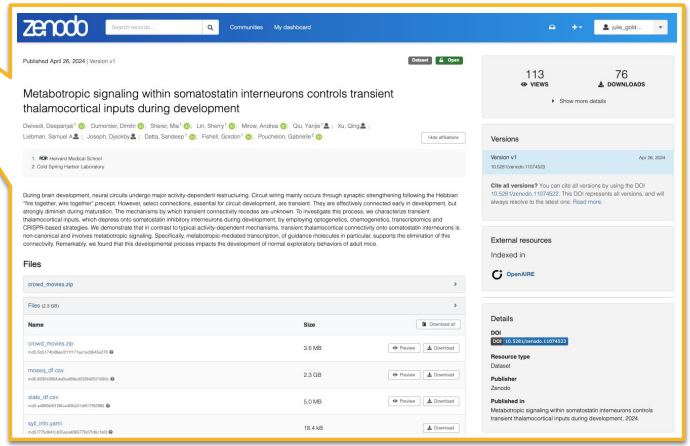
Data availability

All histology and electrophysiology data generated in this study are provided in the Supplementary Information/Source Data files. Publicly available scRNAseq datasets for P2 and P10 cINs utilized in this study are available on Gene Expression Omnibus (GEO) under accession numbers GSE165233 and GSE104156, respectively. The data generated for Motion Sequencing have been deposited in the Zenodo repository https://doi.org/10.5281/zenodo.11074522. Source data are provided with this paper.

Code availability

The Seurat package, which was used for the integration of public developmental datasets and differential gene expression, is open-source and freely available on GitHub (https://github.com/satijalab/seurat) and CRAN. Codes used to analyze the Motion Sequencing data presented in this manuscript are available on GitHub (https://github.com/Pouchelon-Lab/metabotropic_signaling) and provided with the data on Zenodo repository https://doi.org/10.5281/zenodo.11074522.

109. Dwivedi, D. et al. Metabotropic signaling within somatostatin interneurons controls transient thalamocortical inputs during development. *Zenodo* https://doi.org/10.5281/zenodo.11074523 (2024).



Example: Dwivedi, D., Dumontier, D., Sherer, M. et al. Metabotropic signaling within somatostatin interneurons controls transient thalamocortical inputs during development. Nat Commun 15, 5421 (2024). https://doi.org/10.1038/s41467-024-49732-w

Open Researcher and Contributor ID

- ORCID: Provides a persistent digital identifier that distinguishes you from every other researcher and supports automated linkages between you and your professional activities ensuring that your work is recognized
- URI with a 16-digit number that is compatible with the ISO Standard (ISO 27729) or International Standard Name Identifier (ISNI), e.g. https://orcid.org/0000-0001-2345-6789



https://orcid.org

Closing Remarks



Image Source: Taron Egerton & Richard Madden on “Carpool Karaoke” Season 2, Episode 18, March 21, 2019

References & Resources

- Borghi, John, et al. 2018. "Support your data: A research data management guide for researchers." Research Ideas and Outcomes 4: e26439. https://doi.org/10.3897/rio.4.e26439
- Briney, Kristin A., Heather L. Coates, and Abigail Goben. 2020. "Foundational practices of research data management." Research Ideas and Outcomes 6: e56508. https://doi.org/10.3897/rio.6.e56508
- Kathawalla, Ummul-Kiram, Priya Silverstein, and Moin Syed. 2021. "Easing into open science: A guide for graduate students and their advisors." Collabra: Psychology 7 (1): 18684. https://doi.org/10.1525/collabra.18684
- McKiernan, Erin C., et al. 2016. "How open science helps researchers succeed." elife 5: e16800. https://doi.org/10.7554/eLife.16800
- Wilkinson, Mark D., et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18
- Wilson, Greg, et al. 2017. "Good enough practices in scientific computing." PLoS computational biology 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510