Data Management:

The First Step in Reproducible Research

Harvard Chan Bioinformatics Core | Tools for Reproducible Research

August 6, 2024

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

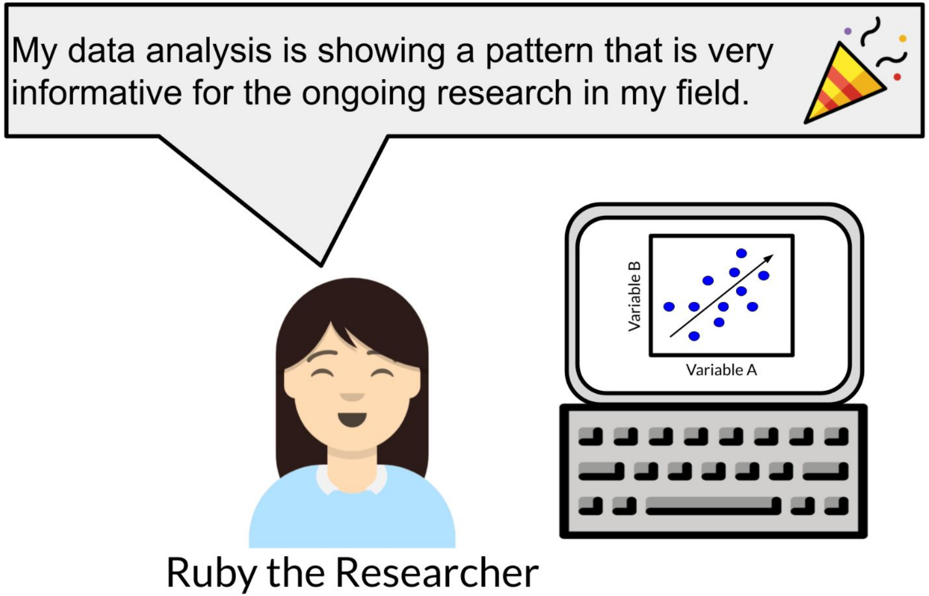
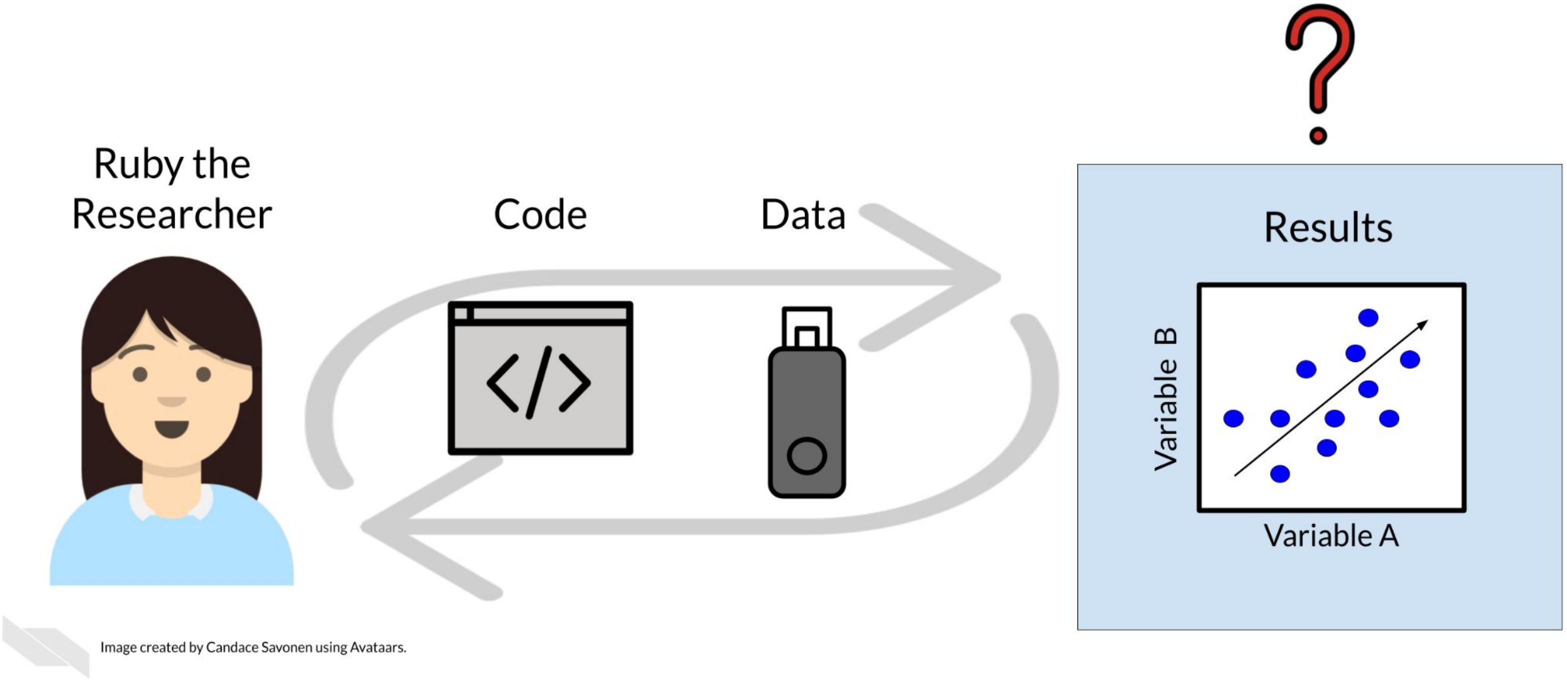

My data analysis is showing a pattern that is very informative for the ongoing research in my field.
Variable B
Variable A
Ruby the Researcher

Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Repeatability



Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Reproducibility

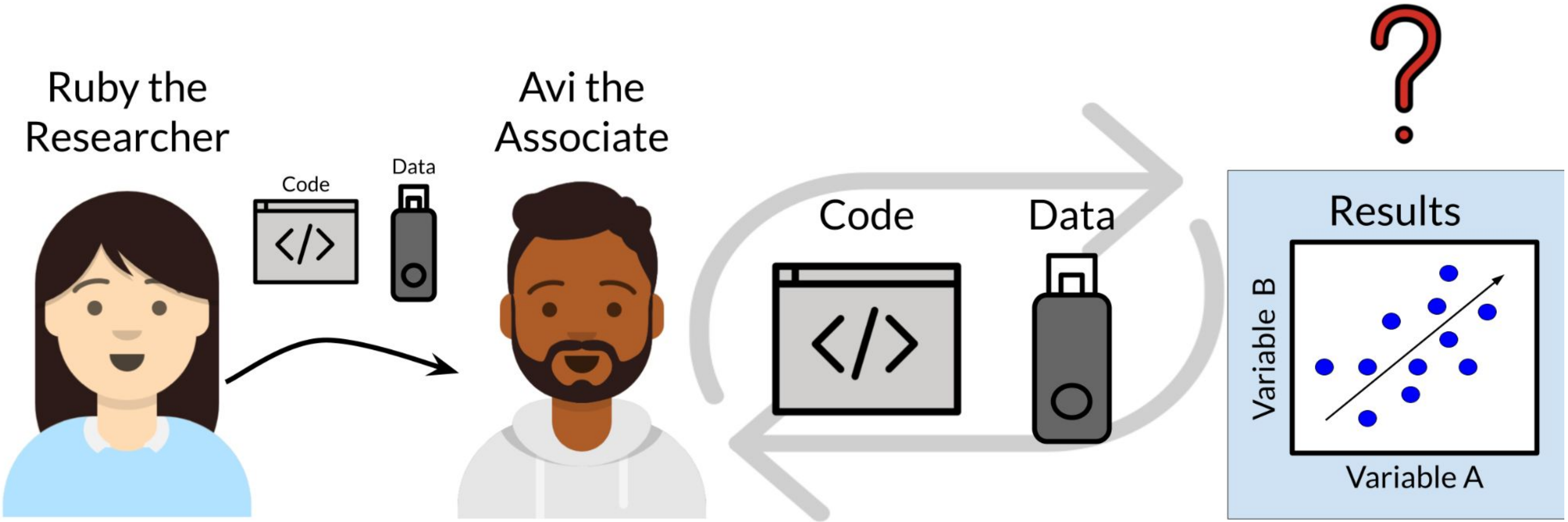
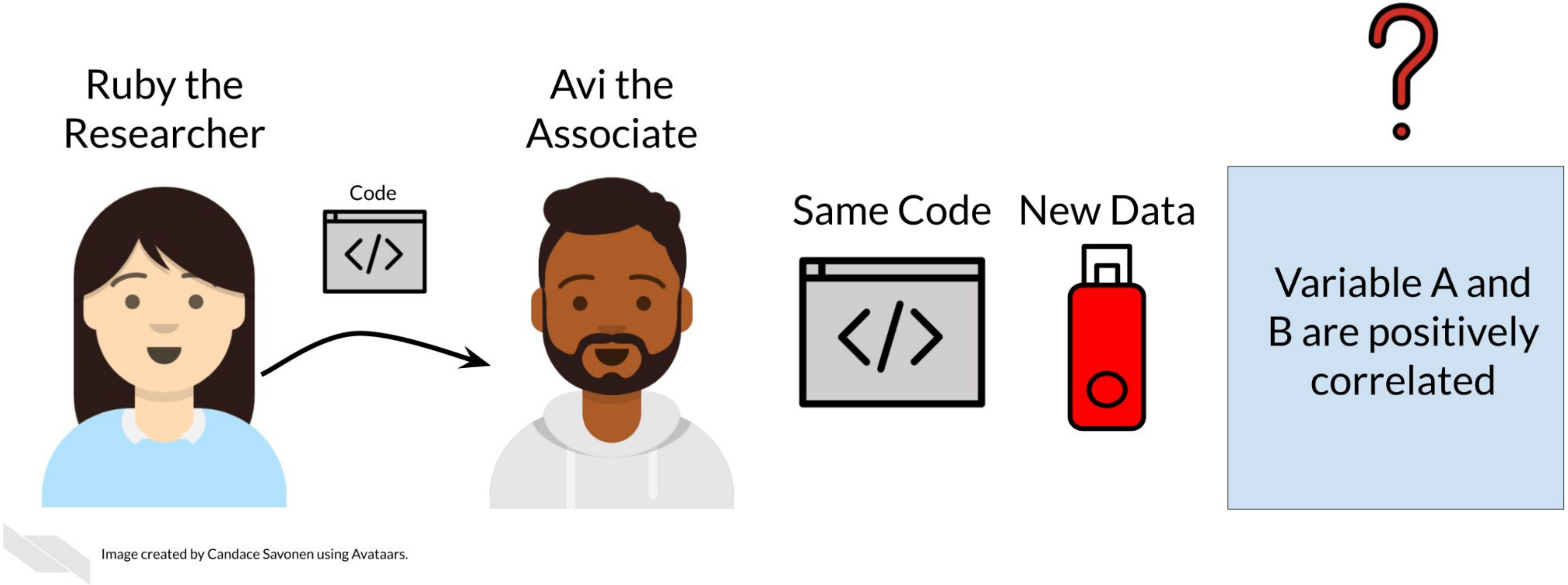


Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Replicability



Ruby the Researcher

Code

Avi the Associate

Same Code

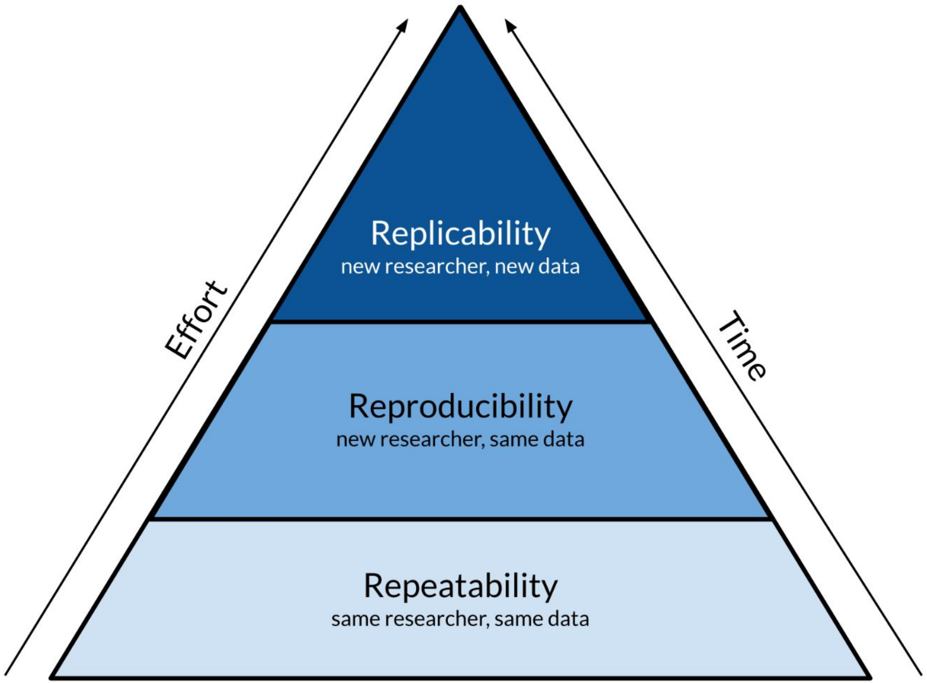
New Data

Variable A and B are positively correlated

Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Research process as a hierarchy



Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics."
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

So, what's the issue?

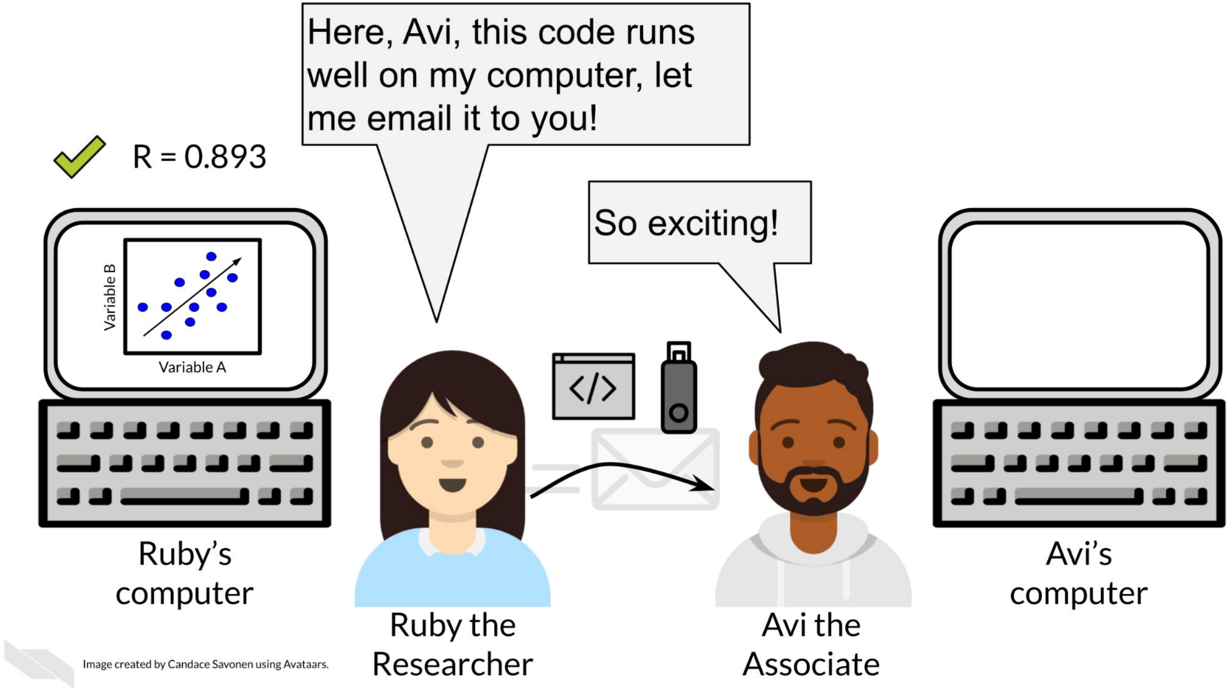
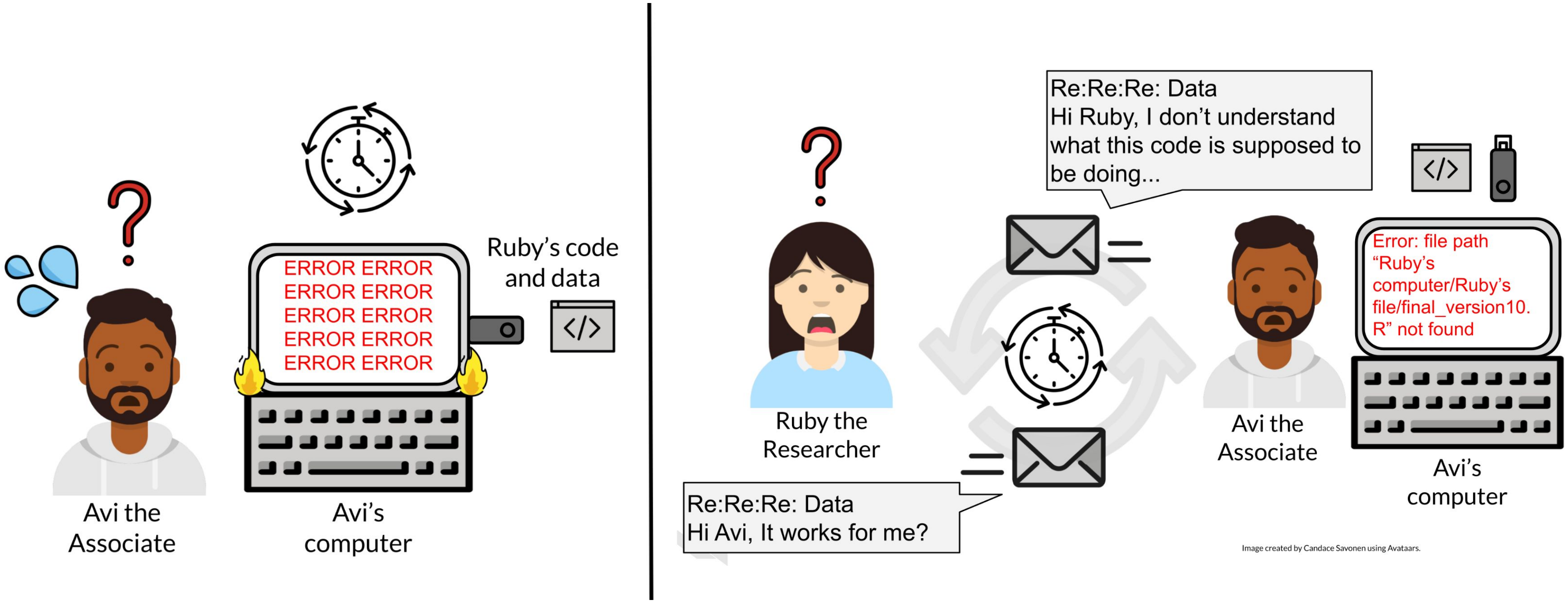


Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics."
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Reproducibility in daily life



ERROR ERROR
ERROR ERROR
ERROR ERROR
ERROR ERROR
ERROR ERROR

Re:Re:Re: Data
Hi Ruby, I don't understand what this code is supposed to be doing...

Re:Re:Re: Data
Hi Avi, It works for me?

Error: file path "Ruby's computer/Ruby's file/final_version10.R" not found

Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics."
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Reproducibility in daily life

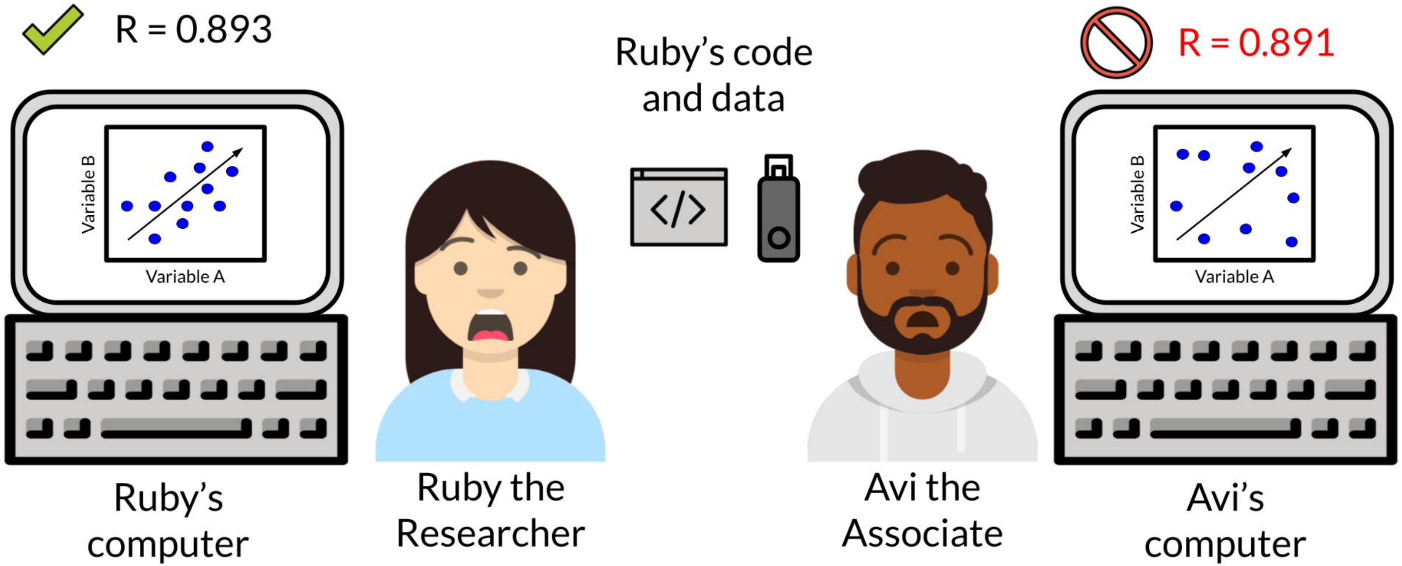
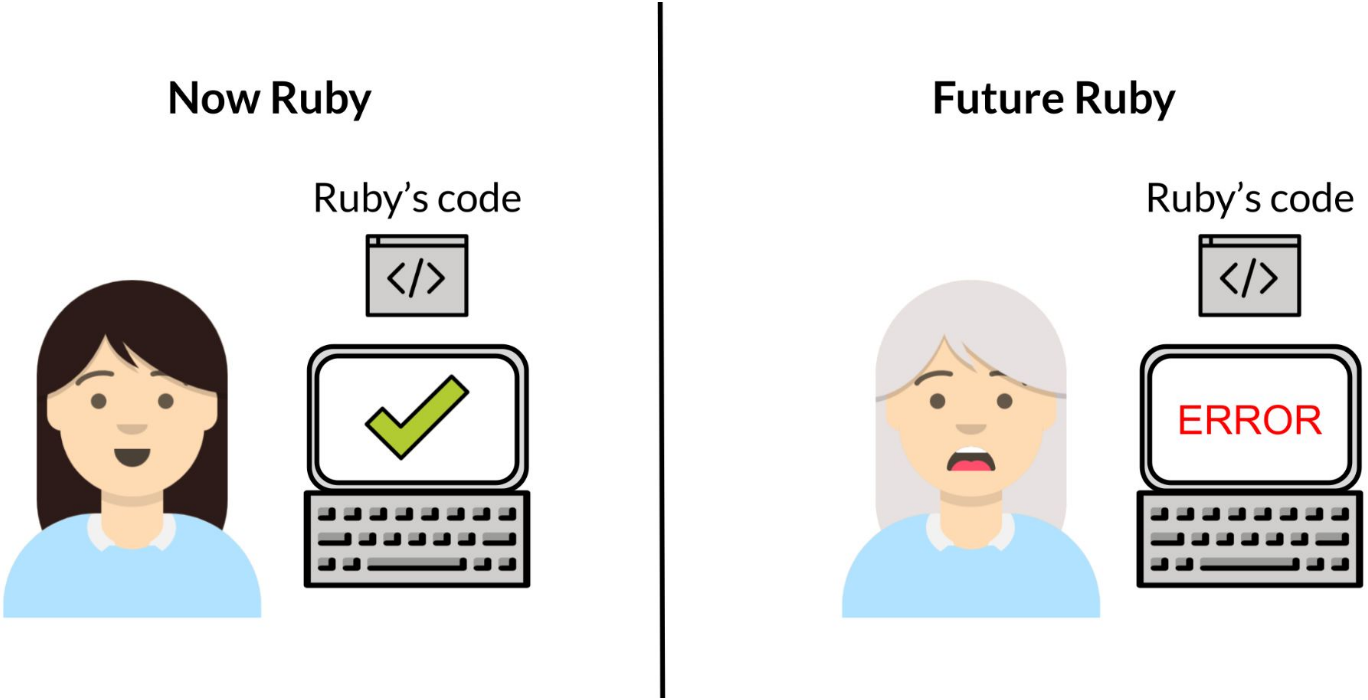


Image created by Candace Savonen using Avataars.

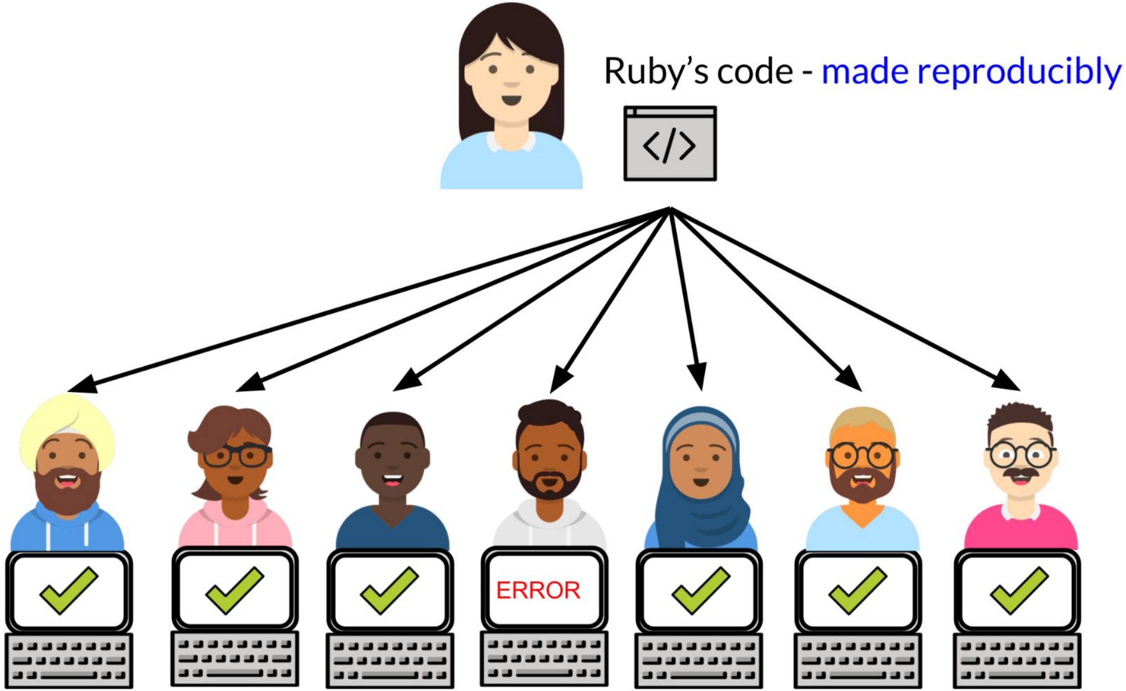
Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Reproducibility in daily life



Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics." https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Reproducibility is worth the effort!



Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.”
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

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

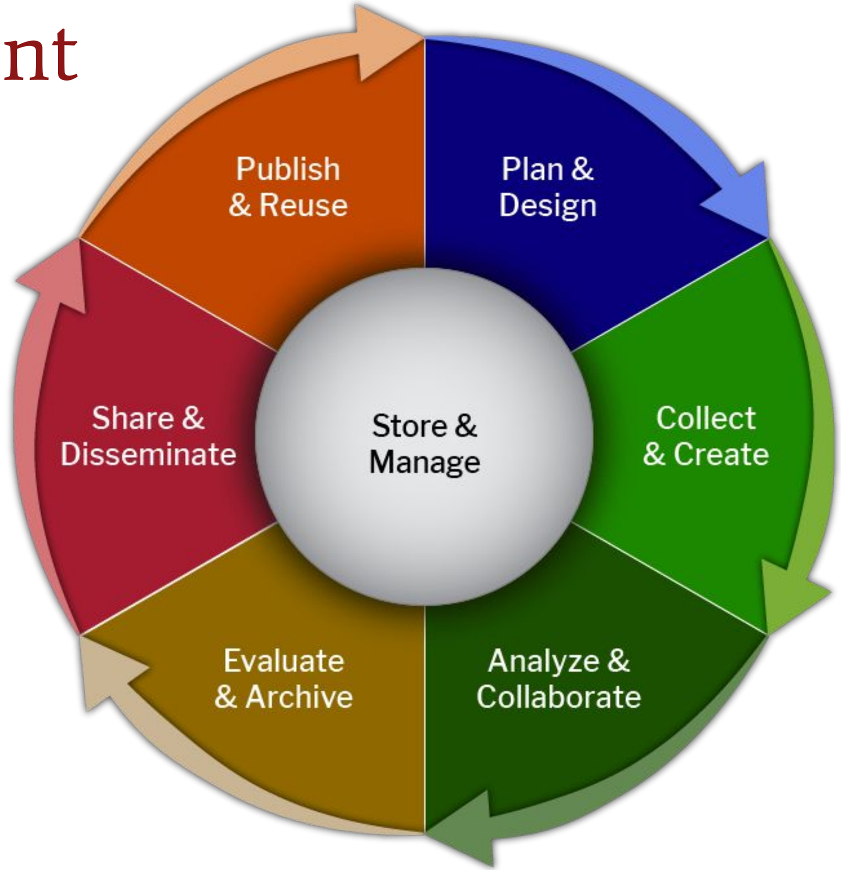
You can't have any sort of reproducibility without good data and project management.

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.

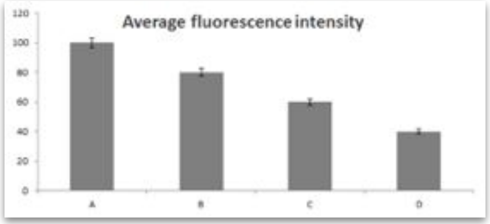
Includes decisions that are agreed to at the beginning of a study and carried out to completion.



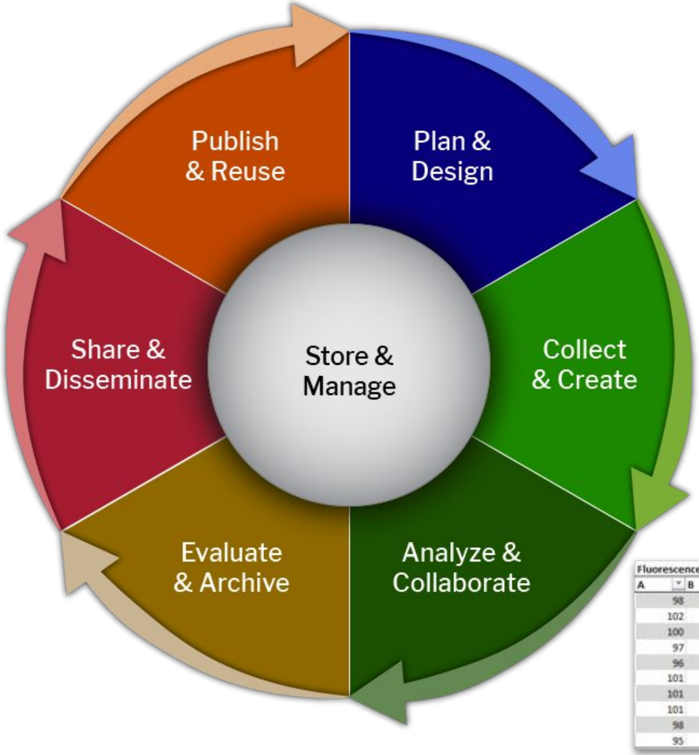
Research Data Lifecycle by LMA RDMWG

Data Through the Research Lifecycle

Final Data

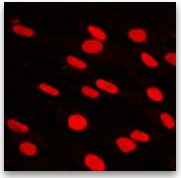


Are the published final data available for validation, reproduction or reuse?

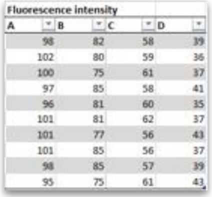


Research Data Lifecycle by LMA RDMWG

What about experimental methods and measurement parameters?



Raw Data



Fluorescence intensity

A	B	C	D
98	82	58	39
102	80	59	36
100	75	61	37
97	85	58	41
96	81	60	35
101	81	62	37
101	77	56	43
101	85	56	37
98	85	57	39
95	75	61	43

Intermediate Data

Data Management Practices for Reproducibility

Organization

- Directory structure
- File naming
- Version control

Documentation

- README File
- Data Dictionary
- Metadata

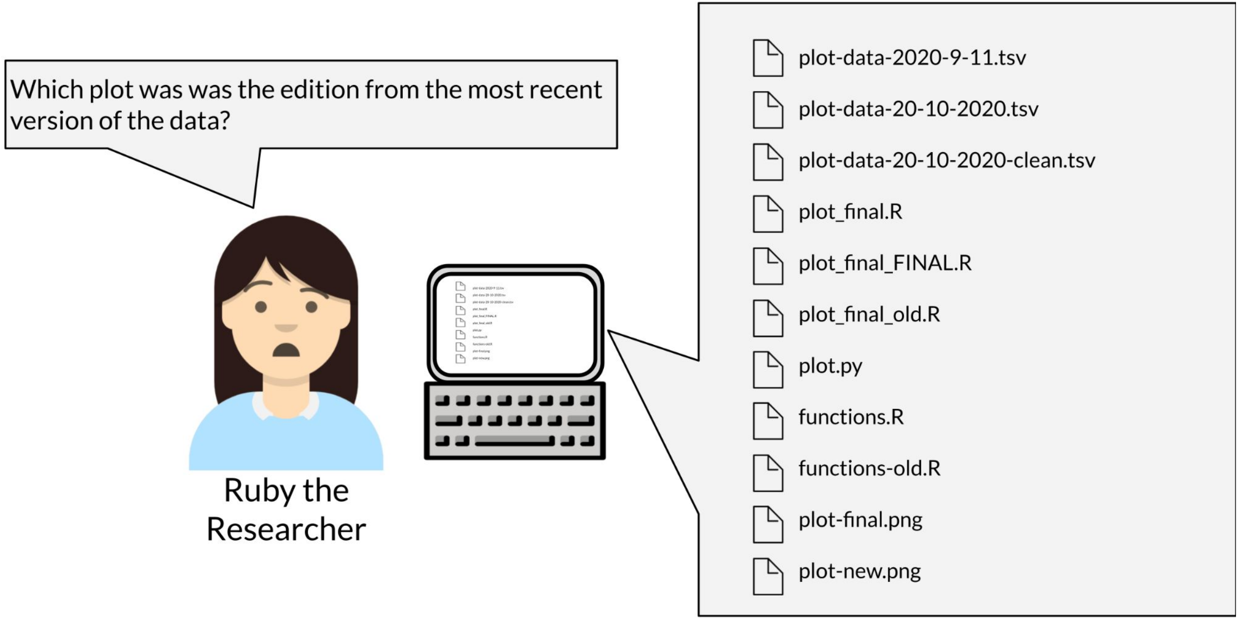
Automation

- Scripts for workflows
- Computing environment
- Dependencies

Dissemination

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

Organization: What to avoid



Which plot was was the edition from the most recent version of the data?

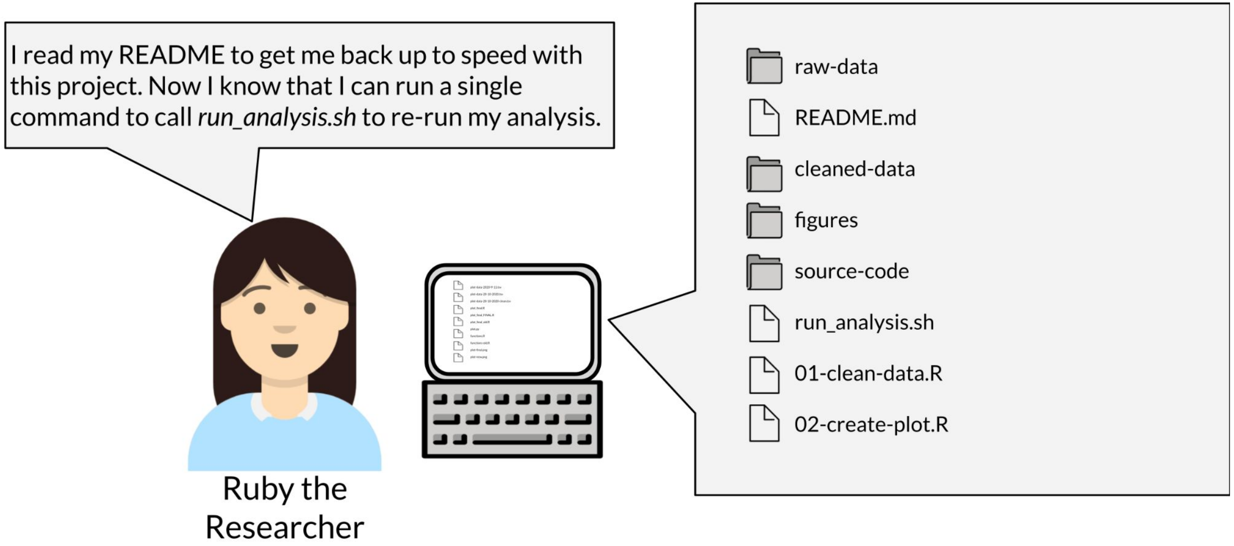
Ruby the Researcher

- plot-data-2020-9-11.tsv
- plot-data-20-10-2020.tsv
- plot-data-20-10-2020-clean.tsv
- plot_final.R
- plot_final_FINAL.R
- plot_final_old.R
- plot.py
- functions.R
- functions-old.R
- plot-final.png
- plot-new.png

Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.”
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Organization: Better practice



I read my README to get me back up to speed with this project. Now I know that I can run a single command to call *run_analysis.sh* to re-run my analysis.

- raw-data
- README.md
- cleaned-data
- figures
- source-code
- run_analysis.sh
- 01-clean-data.R
- 02-create-plot.R

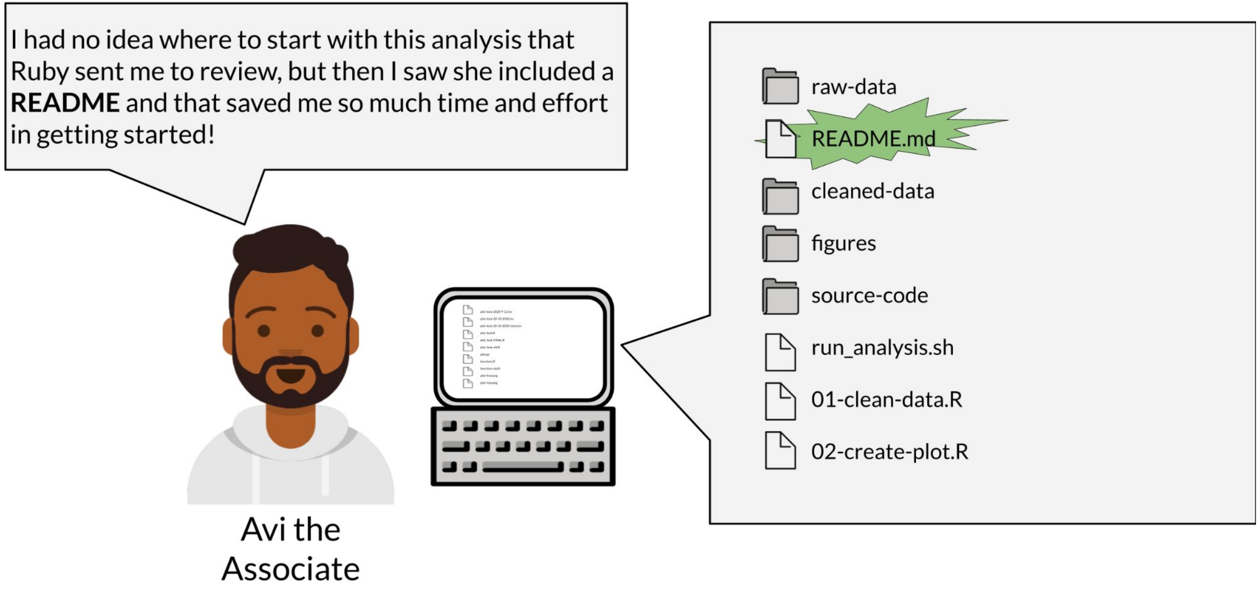
Image created by Candace Savonen using Avataars.

Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics."
https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

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



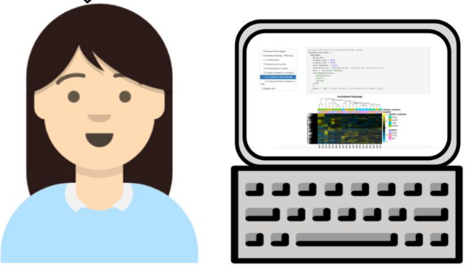
I had no idea where to start with this analysis that Ruby sent me to review, but then I saw she included a **README** and that saved me so much time and effort in getting started!

Image created by Candace Savonen using Avataars.

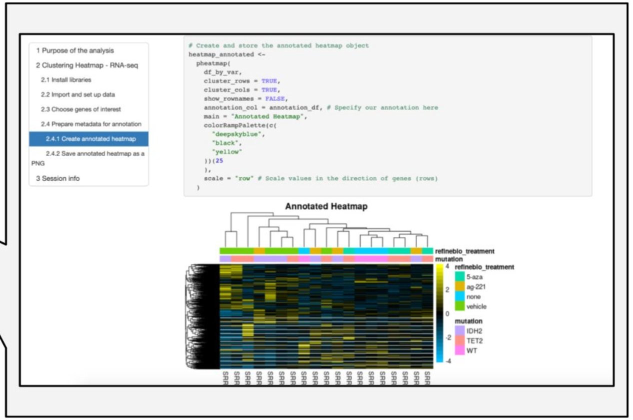
Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Documentation: Good practice

Working from this notebook allows me to interactively develop on my data analysis and write down my thoughts about the process all in one place!



Ruby the Researcher



1 Purpose of the analysis

2 Clustering Heatmap - RNA-seq

2.1 Install libraries

2.2 Import and set up data

2.3 Choose genes of interest

2.4 Prepare metadata for annotation

2.4.1 Create annotated heatmap

2.4.2 Save annotated heatmap as a PNG

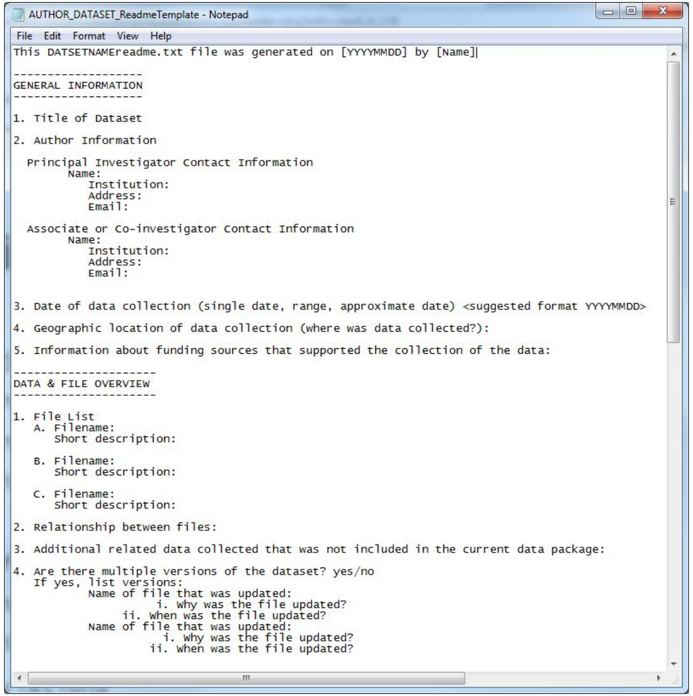
3 Session info

```
# Create and store the annotated heatmap object
heatmap_annotated <-
  pheatmap(
    df_by_var,
    cluster_rows = TRUE,
    cluster_cols = TRUE,
    show_rownames = FALSE,
    annotation_col = annotation_df, # Specify our annotation here
    main = "Annotated Heatmap",
    colorRampPalette(c(
      "deepskyblue",
      "black",
      "yellow"
    ))(25
    ),
    scale = "row" # Scale values in the direction of genes (rows)
  )
```

Image created by Candace Savonen using Avataars.

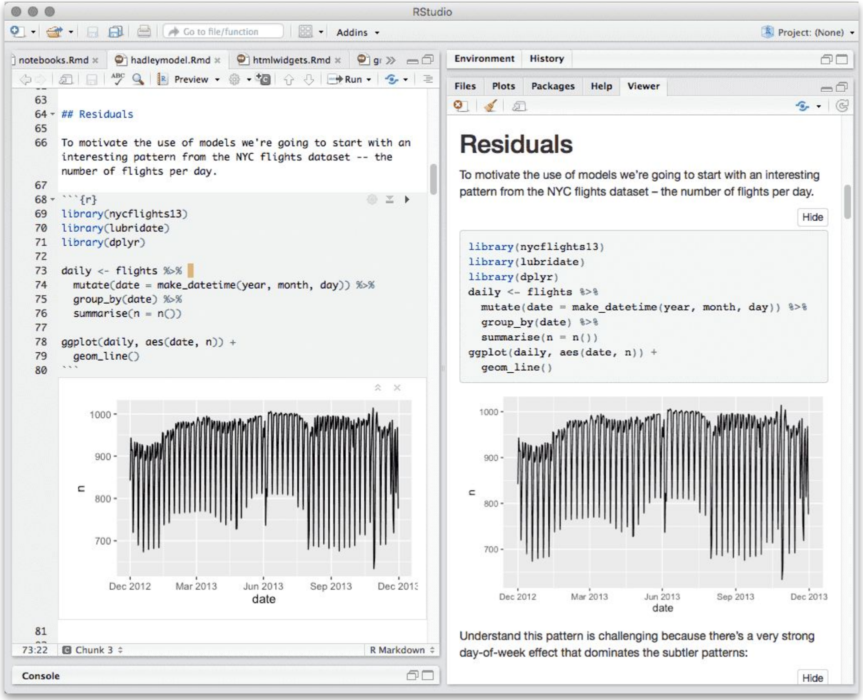
Source: ITCR Training Network (ITN). 2024. “Intro to Reproducibility in Cancer Informatics.” https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Documentation: Useful tools


AUTHOR_DATASET_ReadmeTemplate - Notepad
File Edit Format View Help

```
This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-------------------
GENERAL INFORMATION
-------------------

1. Title of Dataset

2. Author Information

  Principal Investigator Contact Information
        Name:
           Institution:
           Address:
           Email:

  Associate or Co-investigator Contact Information
        Name:
           Institution:
           Address:
           Email:


3. Date of data collection (single date, range, approximate date) <suggested format YYYYMMDD>

4. Geographic location of data collection (where was data collected?):

5. Information about funding sources that supported the collection of the data:


--------------------
DATA & FILE OVERVIEW
--------------------

1. File List
   A. Filename:
      Short description:

   B. Filename:
      Short description:

   C. Filename:
      Short description:

2. Relationship between files:

3. Additional related data collected that was not included in the current data package:

4. Are there multiple versions of the dataset? yes/no
   If yes, list versions:
           Name of file that was updated:
                     i. Why was the file updated?
                ii. When was the file updated?
           Name of file that was updated:
                     i. Why was the file updated?
                   ii. When was the file updated?
```

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



~~~
## Residuals

To motivate the use of models we're going to start with an
interesting pattern from the NYC flights dataset -- the
number of flights per day.

```{r}
library(nycflights13)
library(lubridate)
library(dplyr)

daily <- flights %>%
  mutate(date = make_datetime(year, month, day)) %>%
  group_by(date) %>%
  summarise(n = n())

ggplot(daily, aes(date, n)) +
  geom_line()
```
~~~

Residuals

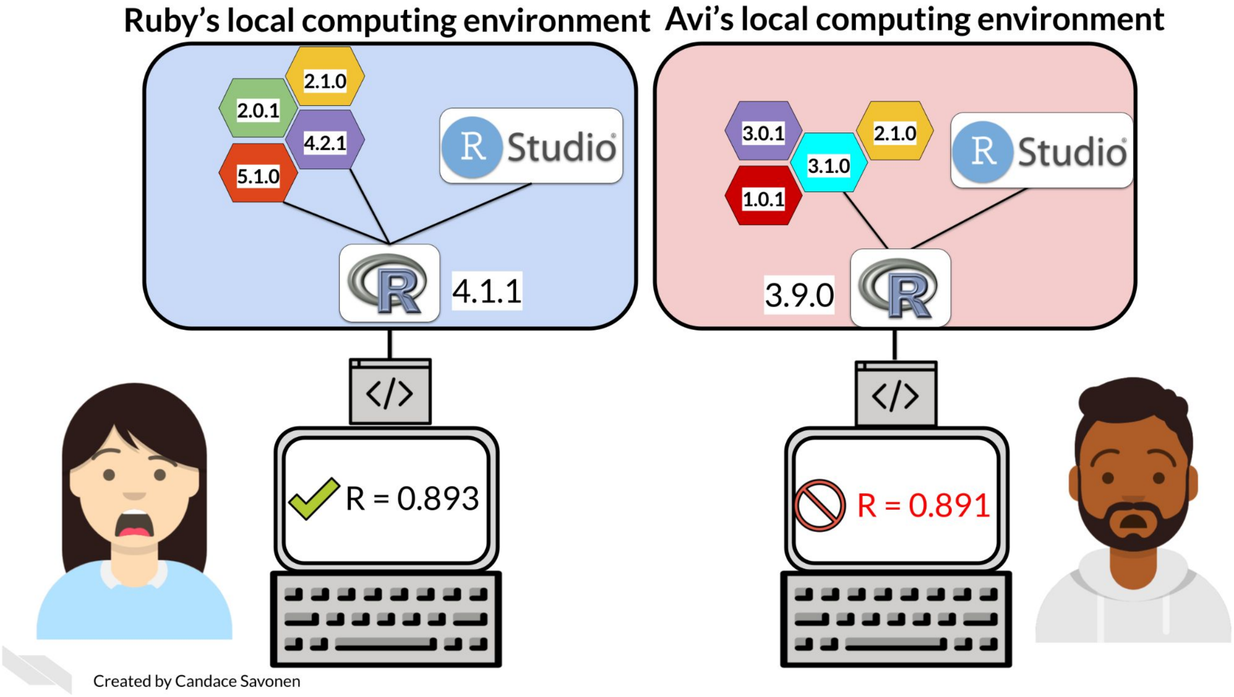
To motivate the use of models we're going to start with an interesting pattern from the NYC flights dataset – the number of flights per day.

```
library(nycflights13)
library(lubridate)
library(dplyr)
daily <- flights %>%
  mutate(date = make_datetime(year, month, day)) %>%
  group_by(date) %>%
  summarise(n = n())
ggplot(daily, aes(date, n)) +
  geom_line()
```

Understand this pattern is challenging because there's a very strong day-of-week effect that dominates the subtler patterns:

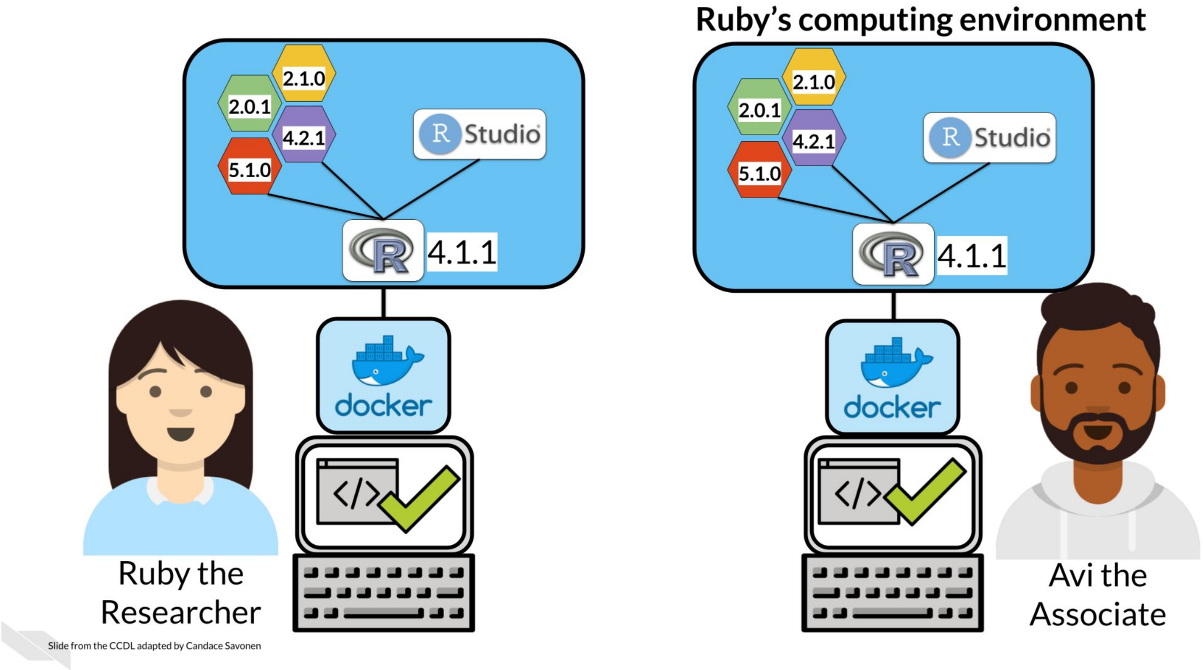
R Markdown: The Definitive Guide:
https://bookdown.org/yihui/rmarkdown/

Automation: What to avoid


Ruby's local computing environment
Avi's local computing environment
2.1.0
2.0.1
4.2.1
5.1.0
R Studio
4.1.1
3.0.1
2.1.0
3.1.0
1.0.1
R Studio
3.9.0
</>
</>
✓ R = 0.893
R = 0.891
Created by Candace Savonen

Source: ITCR Training Network (ITN). 2024. "Advanced Reproducibility in Cancer Informatics." https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics

Automation: Good practice



Source: ITCR Training Network (ITN). 2024. "Advanced Reproducibility in Cancer Informatics."
https://jhudatascience.org/Adv_Reproducibility_in_Cancer_Informatics

Automation: Tips and tricks

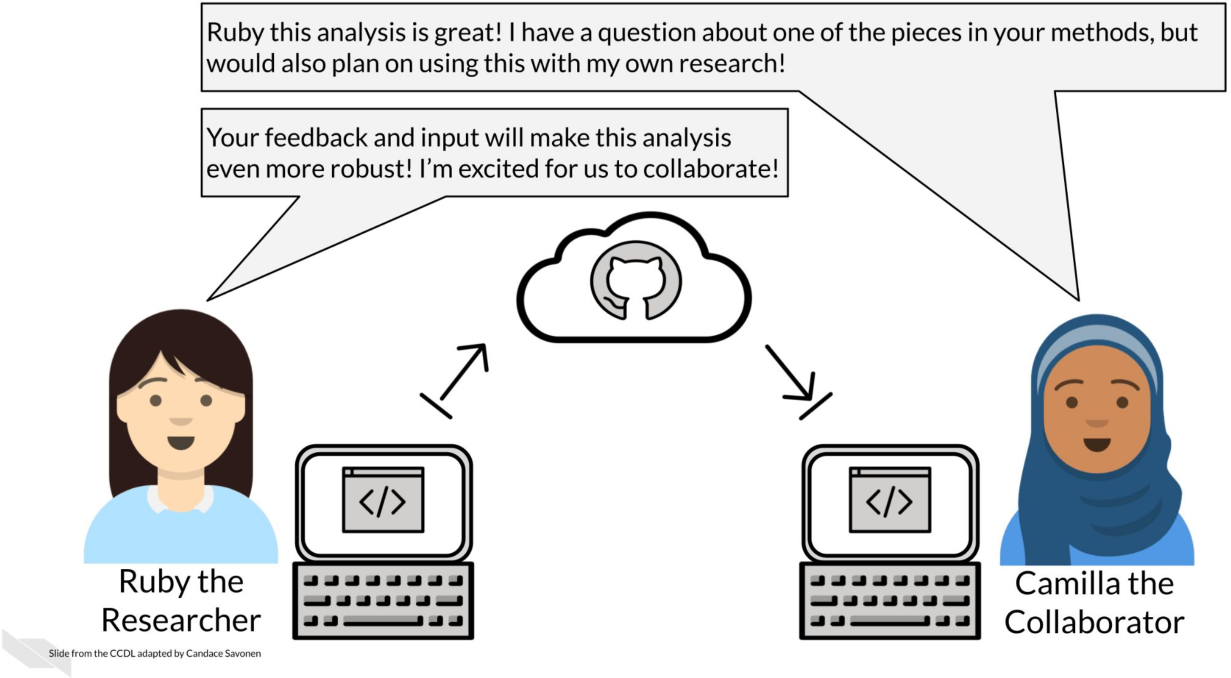
Create a script that can execute all of the various subcomponents of the entire workflow.

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
```

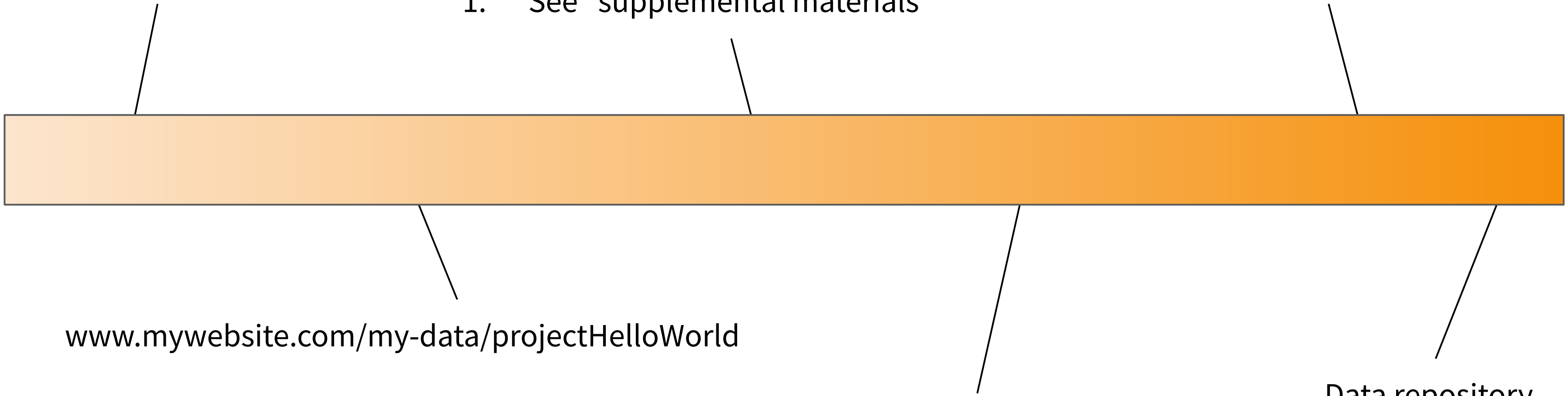
Dissemination: Good practice


Ruby this analysis is great! I have a question about one of the pieces in your methods, but would also plan on using this with my own research!
Your feedback and input will make this analysis even more robust! I'm excited for us to collaborate!
</>
</>
Ruby the Researcher
Camilla the Collaborator

Slide from the CCDL adapted by Candace Savonen

Source: ITCR Training Network (ITN). 2024. "Intro to Reproducibility in Cancer Informatics." https://jhudatascience.org/Reproducibility_in_Cancer_Informatics

Dissemination: Better practice



"Just email me and
I'll send it to you"

www.mywebsite.com/my-data/projectHelloWorld

1. See "supplemental materials"

Dropbox
Box.com
drive.google.com

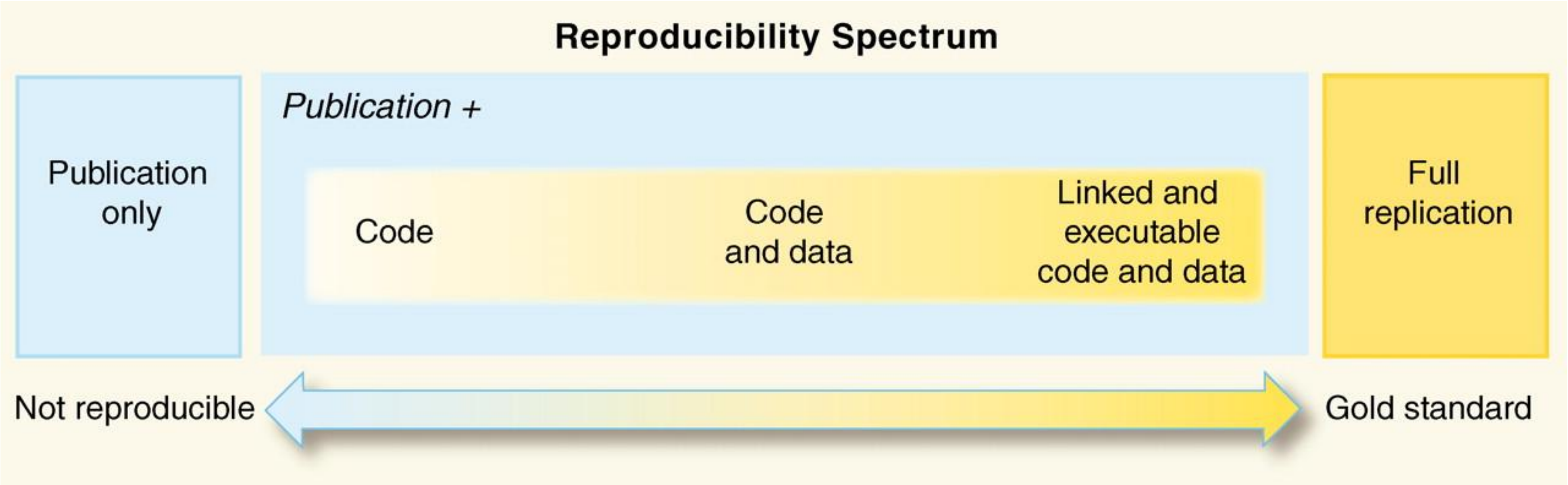
GitHub

Data repository

Dissemination: Useful tools

Disciplinary	General	Software	Methods
db GaP GENOTYPES and PHENOTYPES	The Dataverse® Project	GitHub	protocols.io
GEO Gene Expression Omnibus	OSF	GitLab	natureprotocols
NIMH National Institute of Mental Health	zenodo	Bitbucket	bio-protocol Exchange

Putting it all together!



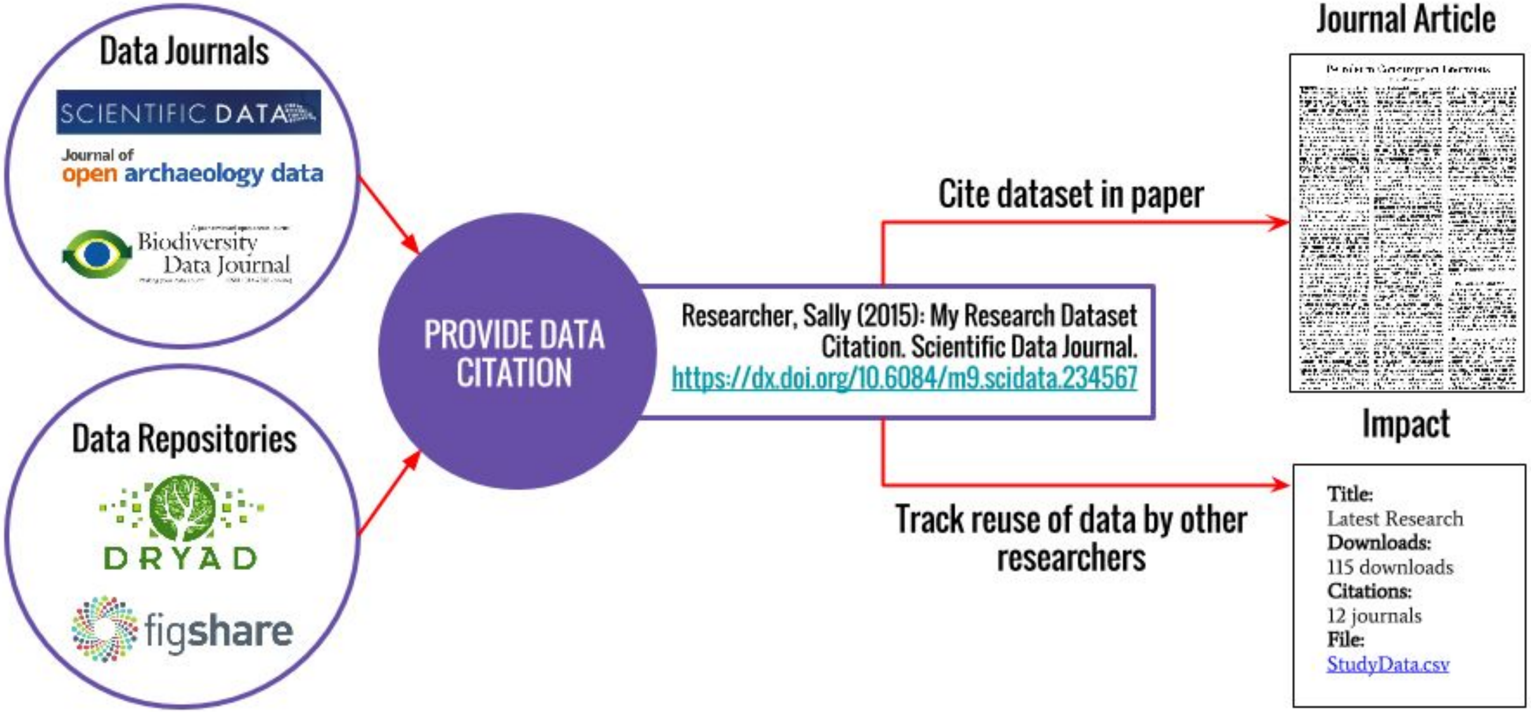
Source: Peng, Roger D. 2011. “Reproducible Research in Computational Science.” *Science* 334 (6060): 1226-1227. https://doi.org/10.1126/science.1213847

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!



Researcher, Sally (2015): My Research Dataset Citation. Scientific Data Journal. https://dx.doi.org/10.6084/m9.scidata.234567

Title: Latest Research
Downloads: 115 downloads
Citations: 12 journals
File: StudyData.csv

Source: Slide courtesy of Vick Steeves. "Building Services Around Reproducibility & Open Scholarship." https://osf.io/pv6ea

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

H.K.
H.K.
U.S.
PhD
Assistant Professor
Marriage
Professor
Professor
Migration
Professor
Retired
iD
0000–1357–2468–0000

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