Everything You Need to Know to Make Your Data Analysis Reproducible

HARVARD LONGWOOD CAMPUS

November 15

Instructor

Julie Goldman
Research Data Services Librarian
Countway Library of Medicine
Julie_Goldman@hms.harvard.edu

Happy to help with data organization, cleaning, and sharing for fostering reproducible workflows and open science!

Book an appointment:
bit.ly/Countway-RDM

Bookmark this website!

https://datamanagement.hms.harvard.edu

Learning Objectives

- Understand the important impact of creating reproducible research
- Establish a reproducible workflow within the context of an example
- Know services and tools available to support reproducible research

What does reproducibility mean?

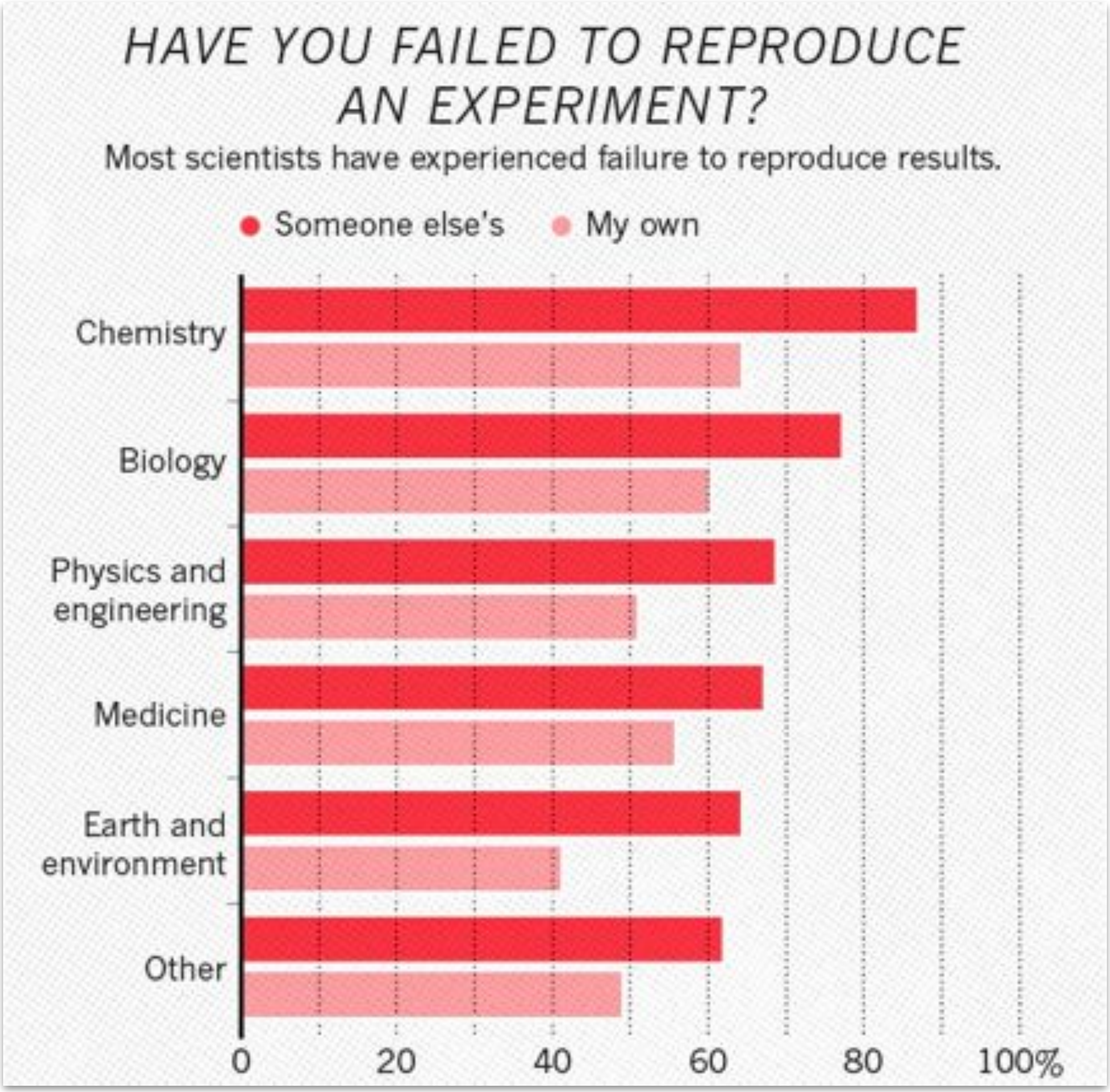
Reproducible Research: Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results.

Replication: A study that arrives at the same scientific findings as another study, collecting new data and completing new analyses.

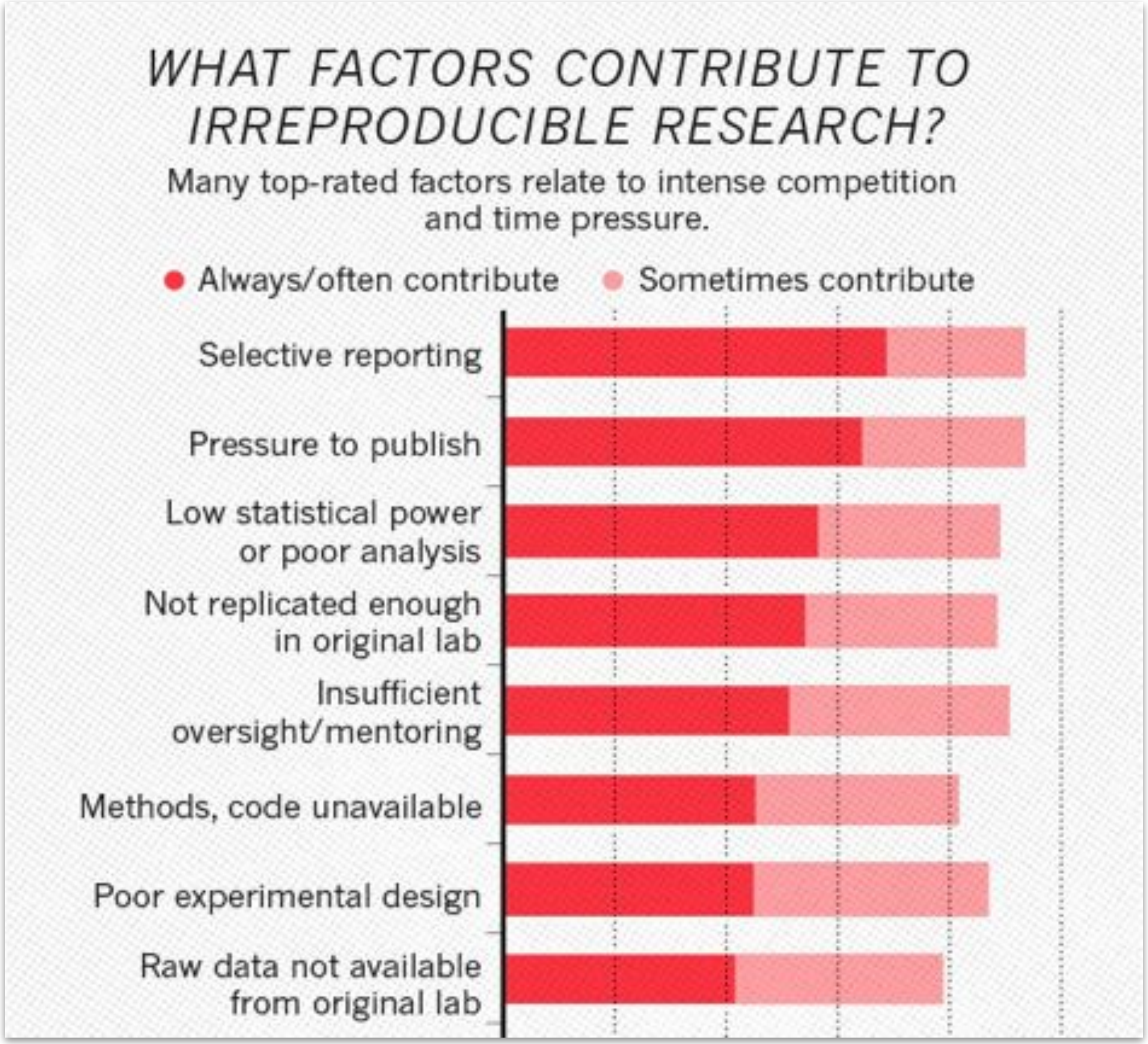
Barba, Lorena A. 2018. "Terminologies for reproducible research." *arXiv preprint* arXiv:1802.03311. https://arxiv.org/abs/1802.03311

Schloss, Patrick D. 2018. "Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research." *MBio* 9(3): e00525-18. http://dx.doi.org/10.1128/mBio.00525-18

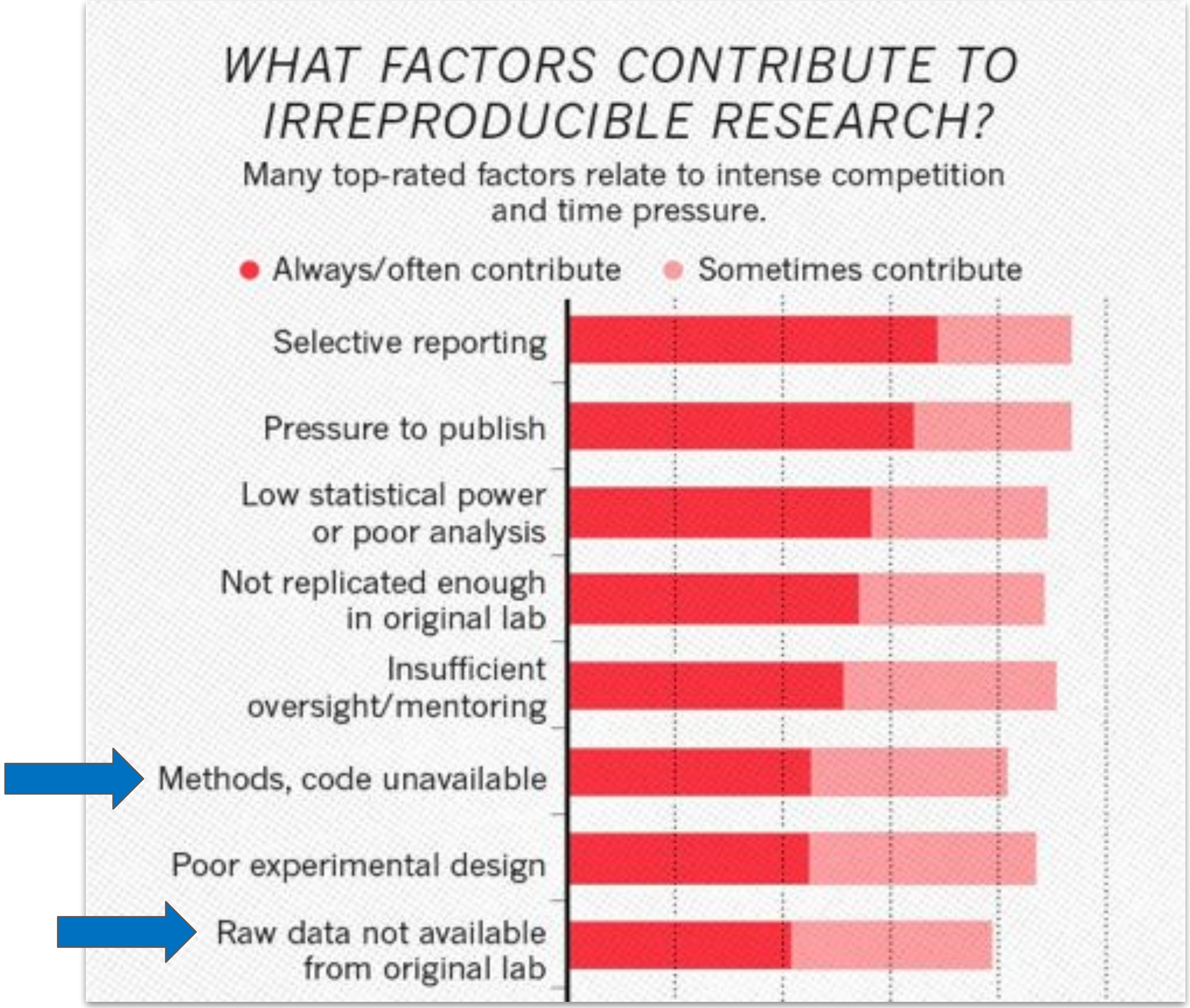
Why does reproducibility matter to ***you***?



Baker, Monya. "1,500 scientists lift the lid on reproducibility." *Nature News* 533(7604): 452-454. https://dx.doi.org/10.1038/533452a



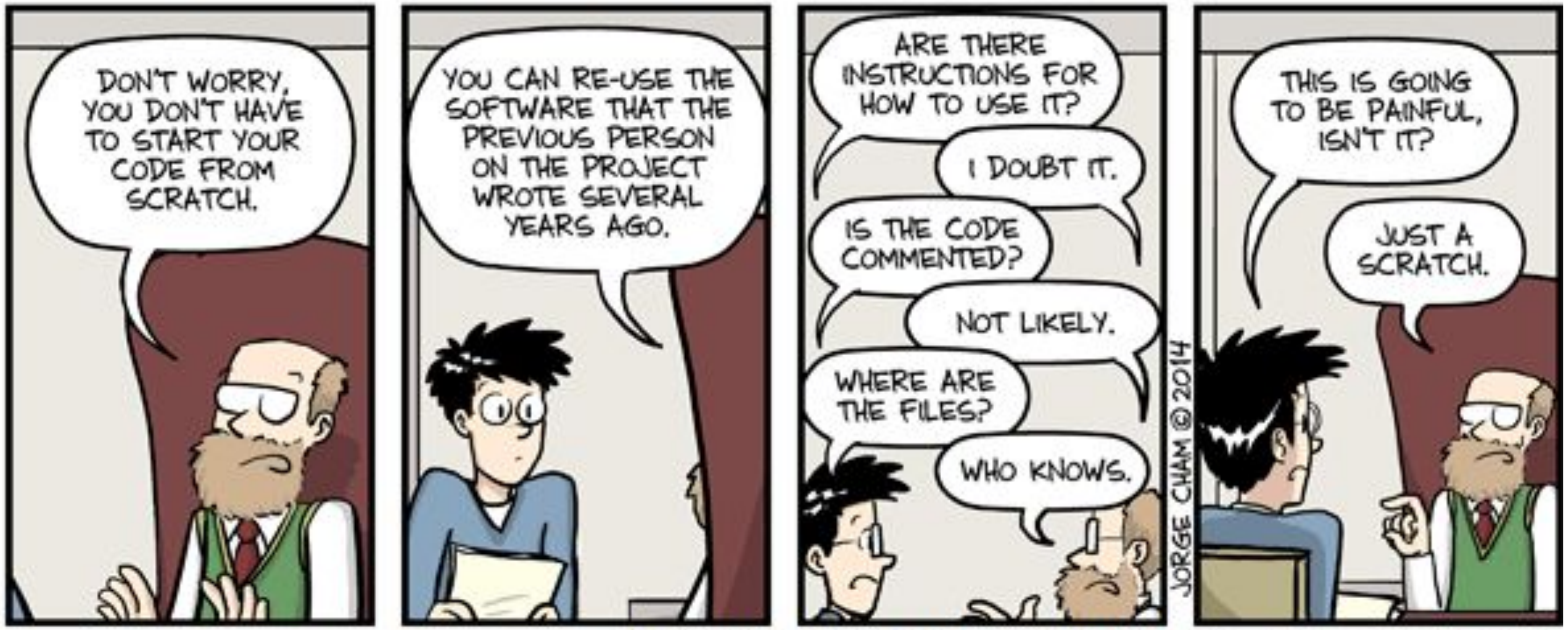
Baker, Monya. "1,500 scientists lift the lid on reproducibility." *Nature News* 533(7604): 452-454. https://dx.doi.org/10.1038/533452a



Baker, Monya. "1,500 scientists lift the lid on reproducibility." *Nature News* 533(7604): 452-454. https://dx.doi.org/10.1038/533452a

What problem are we trying to solve today?

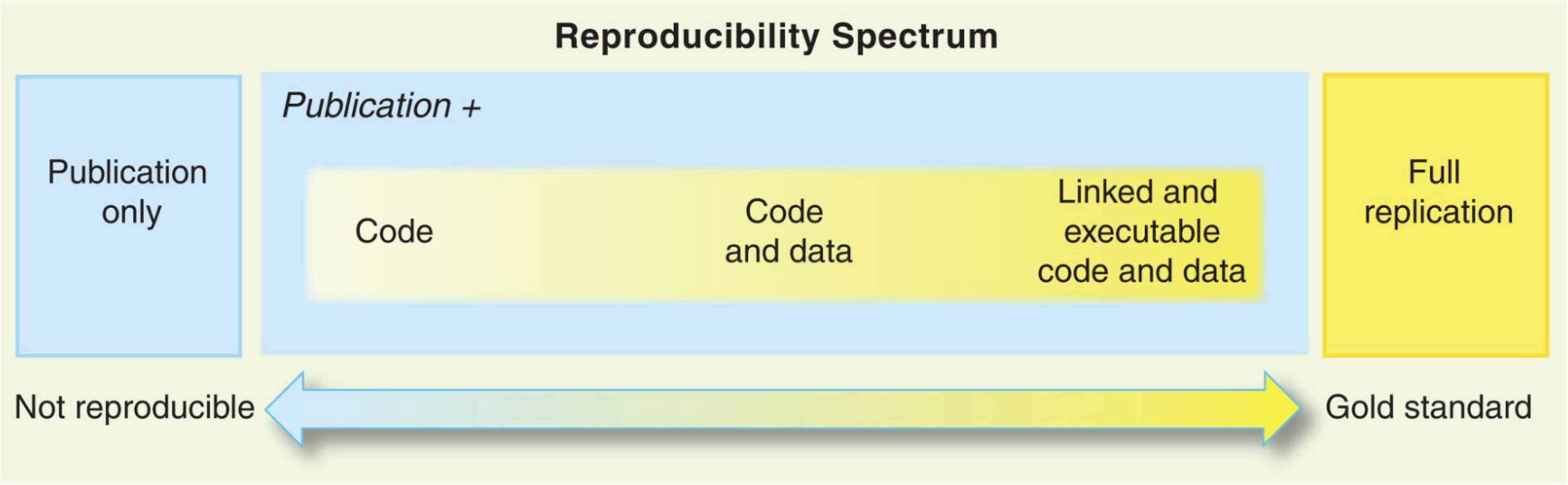
Piled Higher and Deeper *by Jorge Cham*



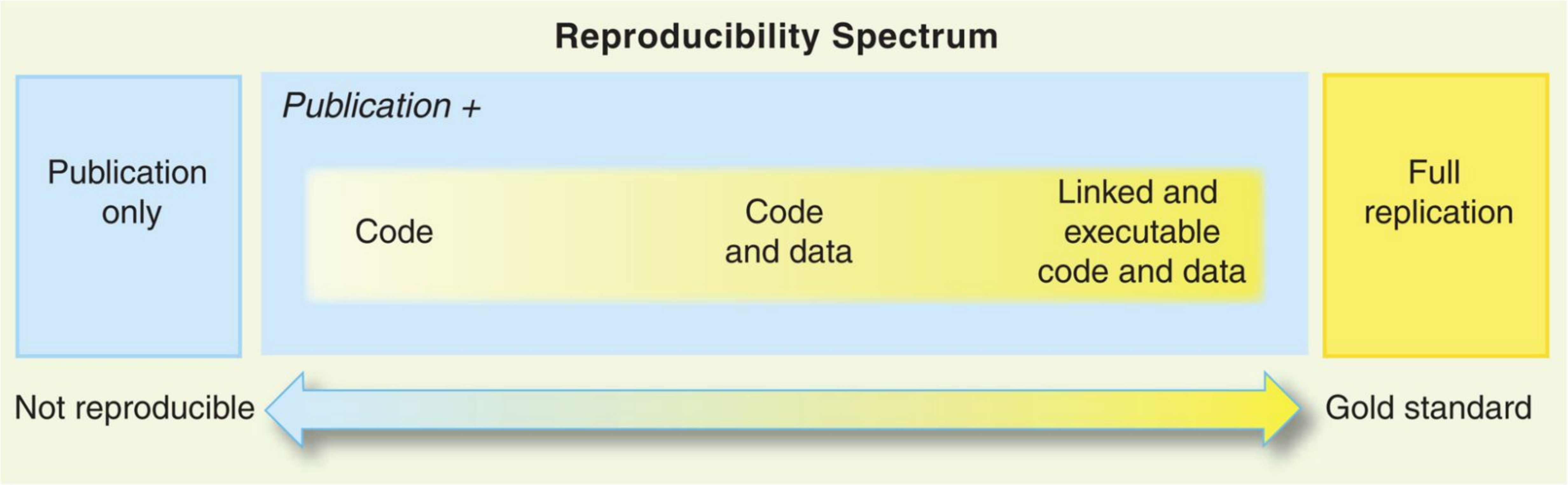
http://phdcomics.com/comics.php?f=1689 title: "Scratch" - originally published 3/12/2014 WWW.PHDCOMICS.COM

Research isn’t being efficiently managed or made reproducible

Much of the time, the workflow & processes aren’t reproducible, the findings (data, code, etc.) aren’t managed efficiently, and as a result, we all suffer.



Peng, Roger D. 2011 "Reproducible research in computational science." *Science* 334(6060): 1226-1227.
https://doi.org/10.1126/science.1213847



The descriptions of the research methods can be independently assessed and the results judged credible (Does not necessarily imply reproducibility.)

Well-documented and fully open code and data allowing others to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

Peng, Roger D. 2011 "Reproducible research in computational science." *Science* 334(6060): 1226-1227. https://doi.org/10.1126/science.1213847

Why Reproducibility

- To build on top of previous work – science is incremental!
- To verify the correctness of results
- To defeat self-deception
- To help newcomers
- To increase impact, visibility and research quality

Nuzzo, Regina. 2015 "How scientists fool themselves–and how they can stop." *Nature News* 526(7572): 182. http://dx.doi.org/10.1038/526182a

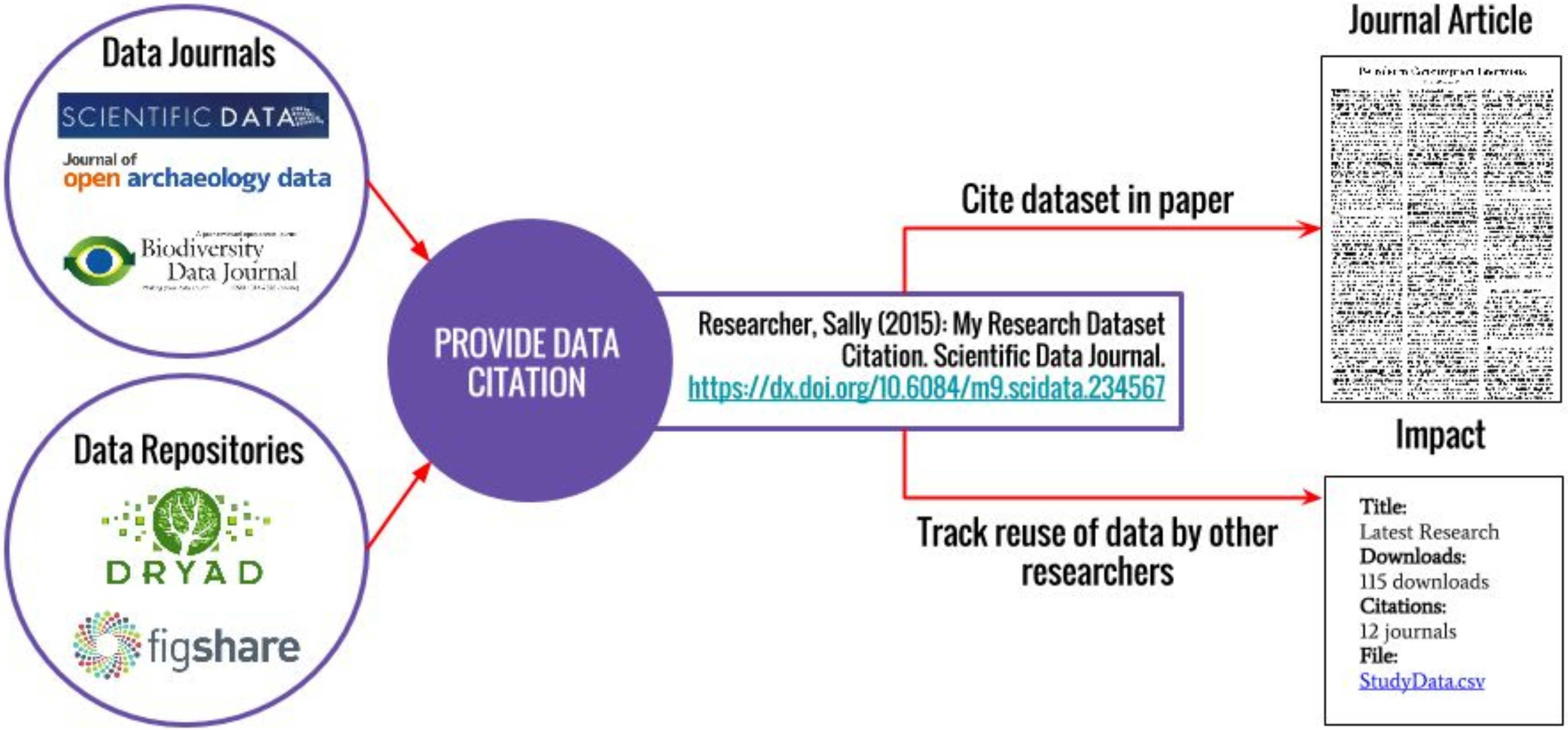
Vandewalle, Patrick, Jelena Kovacevic, and Martin Vetterli. 2009. "Reproducible research in signal processing." *IEEE Signal Processing Magazine* 26(3): 37-47. http://dx.doi.org/10.1109/MSP.2009.932122

Begley, C. Glenn, and Lee M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483(7391): 531. https://doi.org/10.1038/483531a

Why Reproducibility -- think selflessly!

- Others can re-use and extend your work more easily!
  - You can even find interesting collaborations and future research projects out of this
- YOU can re-use and extend your work more easily!
  - Future you is your greatest collaborator
- Newbies to the field can more easily learn the methods by reproducing your work!
  - Your reproducible work is their greatest teacher

Why Reproducibility -- think selfishly!



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

Challenges in Reproducibility

- People make mistakes - and it impacts their research
- It's good to have other people check out your data and analyses - it's like having a copy editor for your data!

- It's *hard* to keep track of what version of what was used
- Software get updates, and these changes can disrupt reproducibility

Incentive Problem

It is a time commitment to get data and code ready to share, and to share it

Reproducibility takes time, and is not always valued by the academic reward structure

"Insufficient time is the main reason why scientists do not make their data and experiment available and reproducible."

Carol Tenopir, Beyond the PDF2 Conference

"77% claim that they do not have time to document and clean up the code."

Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

Pipeline Problem

Technical Obsolescence: Technology changes affect the reproducibility

Normative Dissonance: Ideal values don't always match practice

Reproducibility requires skills that are often not included in most curriculums!

> ***"It would require huge amount of effort to make our code work with the latest versions of these tools."***
>
> Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

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

"Research data management concerns the organization of data, from its entry to the research cycle through the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information."

Whyte, A., Tedds, J. 2011. 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre.
http://www.dcc.ac.uk/resources/briefing-papers

Stages of the Research Data Management Lifecycle

Managing the way data is collected, processed, analyzed, preserved, and published for greater reuse by the community and the original researcher.

PLAN	COLLECT, GENERATE & STORE	CLEAN, ANALYZE & VISUALIZE	PUBLISH & SHARE	ARCHIVE & PRESERVE	REUSE
Plan for research data needs	Acquire, organize & store data	Process data for current use	Organize, describe & share data	Appraise, preserve & steward data	Discover & reuse data

Data and Project Management

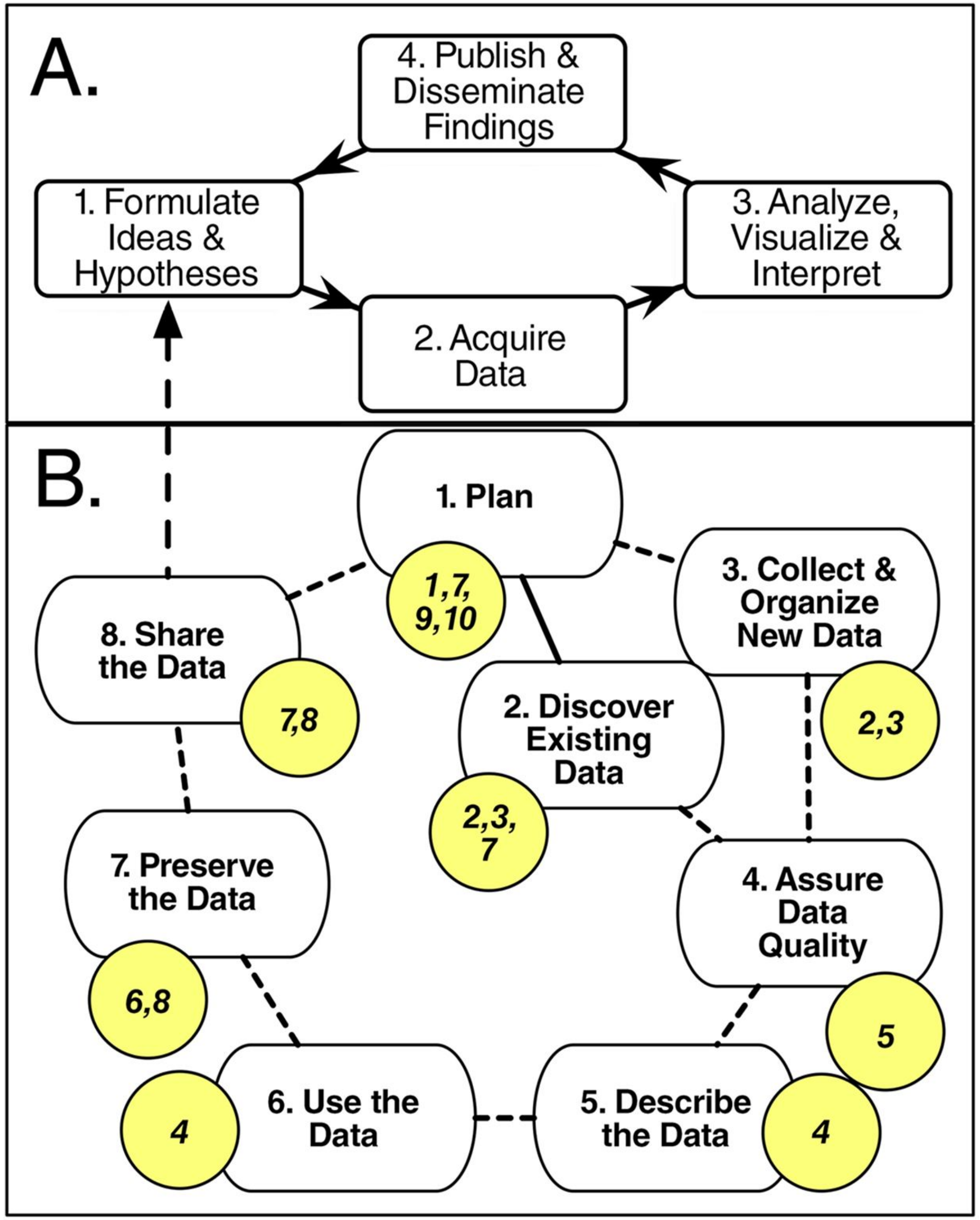
Document the activities for the entire lifecycle. Create a Data Management Plan including sponsorship requirements, realistic budget, assigned responsibilities, all the data to be collected, and each of these topics!

- Consider what do you want to get out of managing your data
- Figure out your criteria for keeping data
- Think about where you want your data
- Consider the metadata you want to collect to document your datasets

https://datamanagement.hms.harvard.edu/planning-overview

Data Management Plan

1. Determine the Research Sponsor Requirements
2. Identify the Data to Be Collected
3. Define How the Data Will Be Organized
4. Explain How the Data Will Be Documented
5. Describe How Data Quality Will Be Assured
6. Present a Sound Data Storage and Preservation Strategy
7. Define the Project's Data Policies
8. Describe How the Data Will Be Disseminated
9. Assign Roles and Responsibilities
10. Prepare a Realistic Budget



Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput Biol* 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Key Practices for Reproducibility

- **Organization**: Organize your projects so that you don’t have a single folder with hundreds of files and use tools to your advantage
- **Documentation**: Clearly separate, label, and document all data, files, and keep track of operations that occur on data and files using version control
- **Automation**: Design a workflow as a sequence of small steps that are glued together, with intermediate outputs from one step feeding into the next step as inputs and use scripting to create automated data analyses
- **Dissemination**: Publishing is not the end of your analysis, rather it is a way station towards your future research and the future research of others

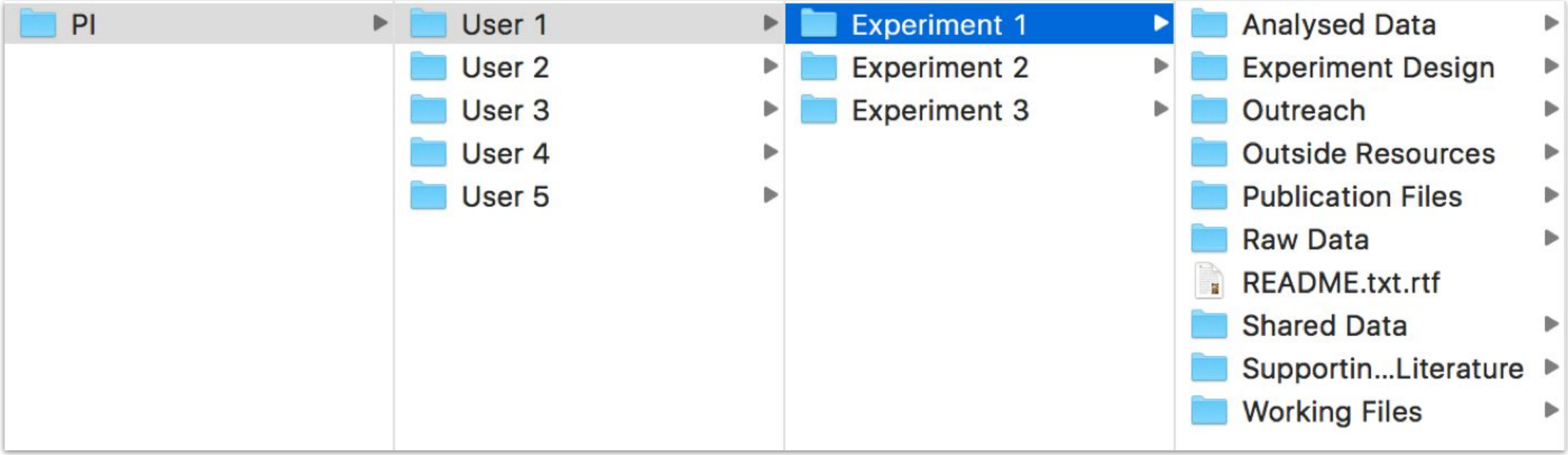
Project Organization

- Put each project in its own directory, which is named after the project
- Put text documents associated with the project in the doc folder
- Put raw data and metadata in the data folder, and files generated during cleanup and analysis in a results folder
- Put source for the project’s scripts and programs in the src folder
- Name all files to reflect their content or function
  - Do not use special characters (!@#$%^*) or spaces
  - Use underscores or dashes, A-Z, and numbers

Example Project Structure

Create a directory structure for output files ***before*** running analysis workflow

- Have README.txt files in higher level directories briefly describing their contents
- Have log files for each tool documenting the versions/parameters used



https://datamanagement.hms.harvard.edu/directory-structure

Project Documentation

LITERALLY EVERYTHING

IF IT HAPPENS DURING YOUR PROJECT...

CHANCES ARE YOU NEED TO

Project Documentation

LITERALLY EVERYTHING

IF IT HAPPENS DURING YOUR PROJECT…

CHANCES ARE YOU NEED TO

DOCUMENT

Data Needs Documentation

Methodology

We are collecting data from 30 women ages 18-25 about their sexual histories through individual interviews.

We will analyze this data using XYZ software and XYZ analytical framework.

take note of changes to this as the project continues

Data Collection

We will use the Open Science Framework to document our data collection process.

“Subject CYZ was interviewed in my office at Harvard Medical School from 1-3pm. The recording file is located in 2018/PROJECT/INTERVIEWS”

Variable Names

Variable Name: **employ_dev**

Description: A derived variable based on the percentage of a given economic development area employed in full time work. Expressed as the value of the variable **employ** divided by the number of work-eligible adults resident in that district as listed in the 1980 census.

Data Needs Documentation

Methodology

We are collecting data from 30 women ages 18-25 about their sexual histories through individual interviews.

We will analyze this data using XYZ software and XYZ analytical framework.

take note of changes to this as the project continues

DOCUMENT

Data Collection

We will use the Open Science Framework to document our data collection process.

“Subject CYZ was interviewed in my office at Harvard Medical School from 1-3pm. The recording file is located in 2018/PROJECT/INTERVIEWS”

DOCUMENT

Variable Names

<u>Variable Name</u>: **employ_dev**

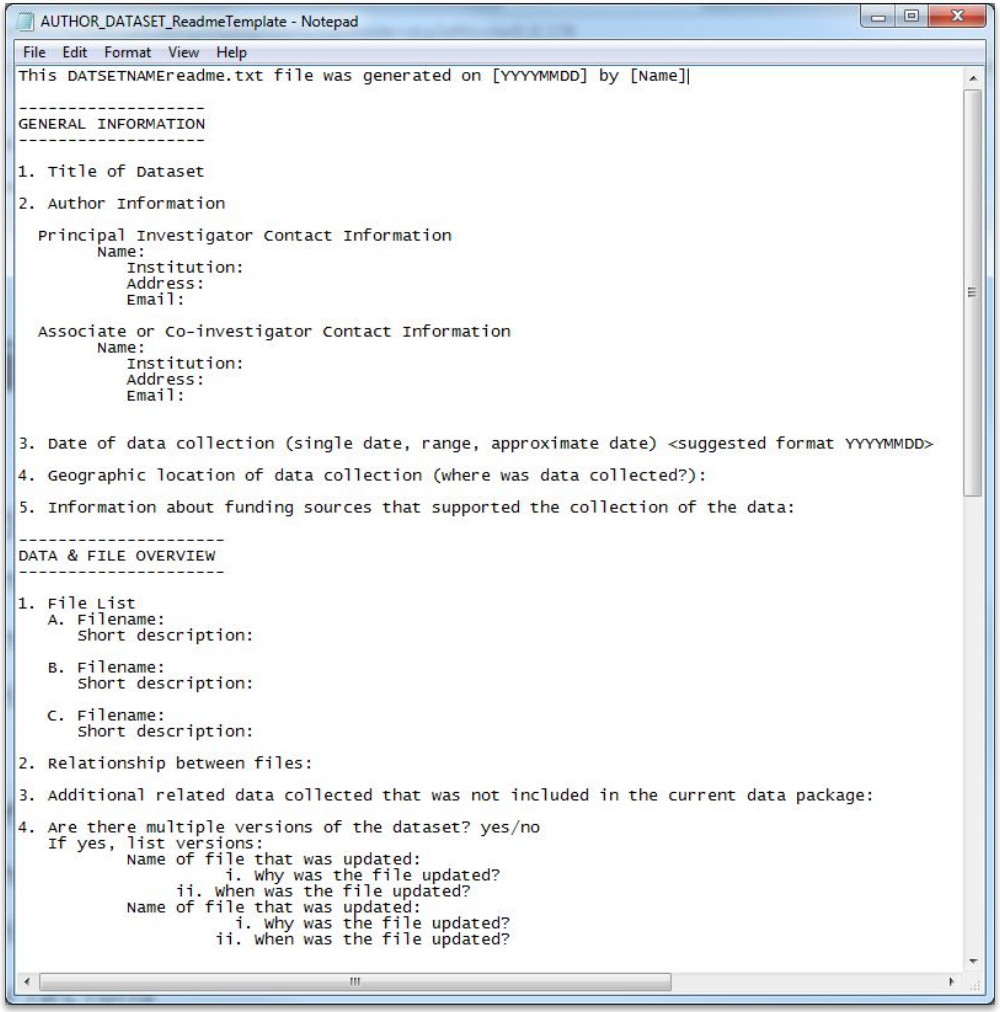
<u>Description</u>: A derived variable based on the percentage of a given economic development area employed in full time work. Expressed as the value of the variable **employ** divided by the number of work-eligible adults resident in that district as listed in the 1980 census.

DOCUMENT

Recipe for Reproducibility → README

1. **Quality Assurance:** Collecting and Validating Data
2. **Describing Data & Providing Documentation:** Contextual information and details about data sets needed for discovery, access, use, and reuse
3. **Code & scripts:** Code associated with analyses and results
4. **Deposit & Disseminating Data (and code):** Submission to open data repositories

Documentation with README.txt



```
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?
```

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

Vital for Every Project: Title, Authors, Description, Date, License

Vital for Programs: Language Versions, Dependencies, Dependency Versions, Git Commit, proper version number, How to Install

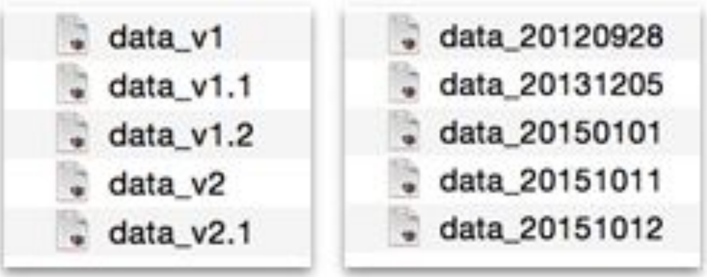
Vital for Larger Programs: Tests, How to run Tests, Run Times Under Commonly Used Platforms, Sample Input and Output Data, Sample Run Usage

Project Version Control

- Version control is used to capture a snapshot of all of a project's files at any moment in time, allowing a researchers to easily review the history of the project and to manage future changes
  - Provides a means of documenting and tracking changes to project files in a systematic and transparent manner
  - Enables seamless collaboration so many people can work on a file at once
  - Helps with reverting to previous (working) versions
- Record changes to scripts (additions/deletions/replacements)
  - What was changed?
  - Who is responsible?
  - When did it happen?

Version Control

Basic – file names



Intermediate – built-in software capabilities

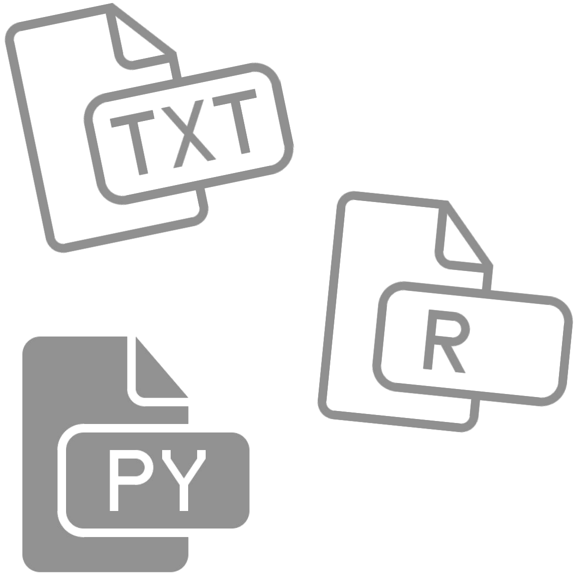
Advanced – version control software

Version Control with Git

- Git is a revision control system, a program to manage your source code history. It is strictly a command-line tool.
- The purpose of Git is to manage a project, or a set of files, as it changes over time. Git stores this information in a data structure called a repository.
- A Git repository contains, among other things, the following:
    - Snapshots of your files (source code, text, etc.)
    - References to these snapshots, called heads

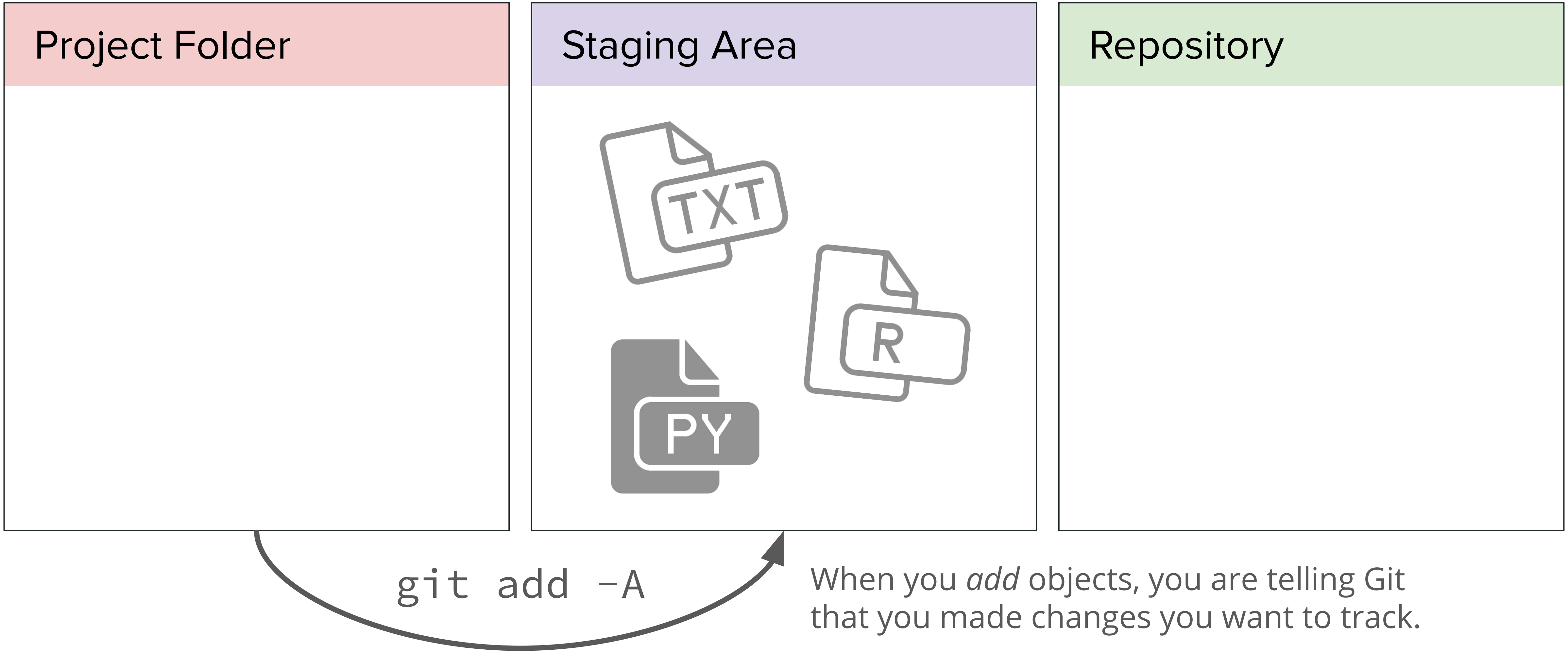
Git Example

* changes not tracked *

Project Folder	Staging Area	Repository
		

Git Example

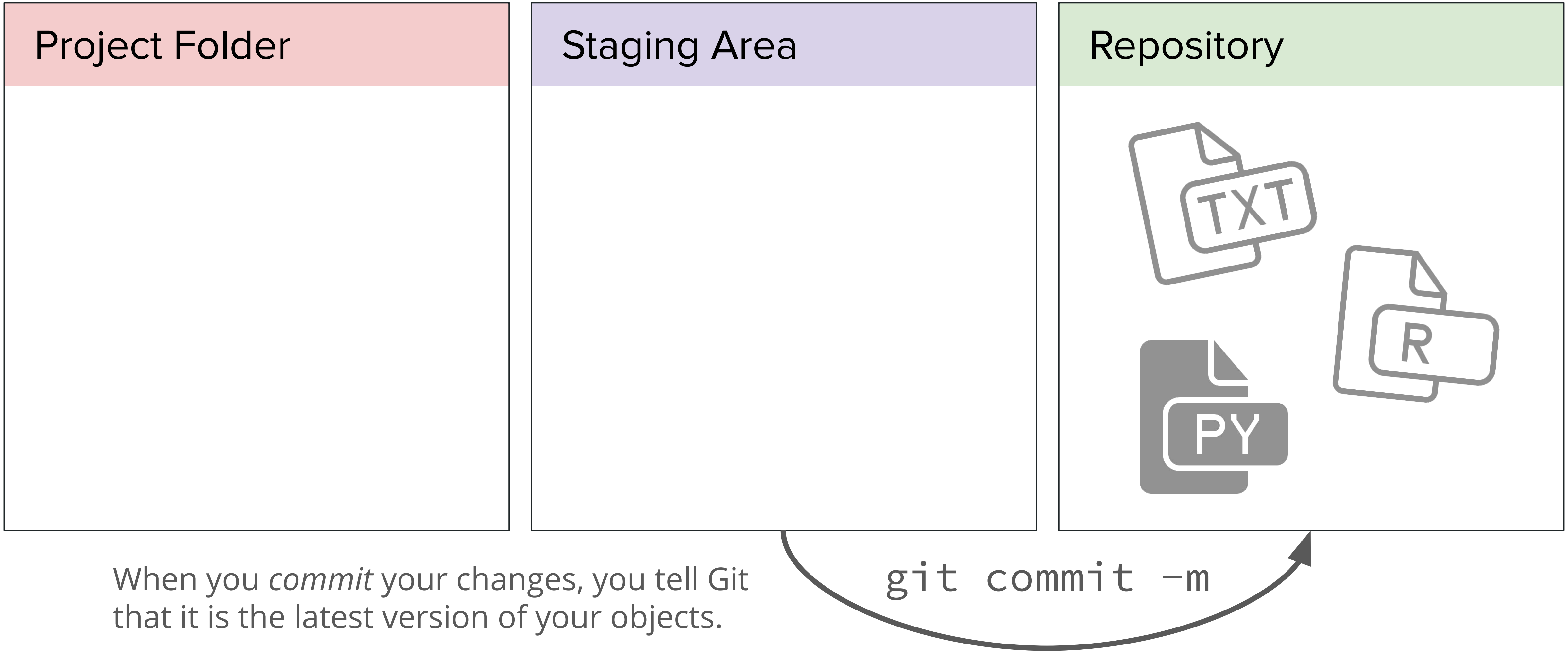
* changes noted but no new version created *



When you *add* objects, you are telling Git that you made changes you want to track.

Git Example

* changes become the new version *



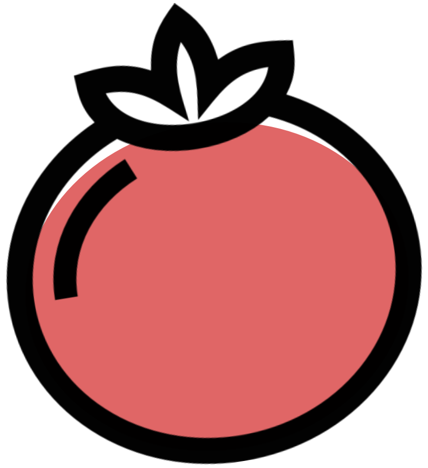
When you *commit* your changes, you tell Git that it is the latest version of your objects.

Reproducible Research Workflow

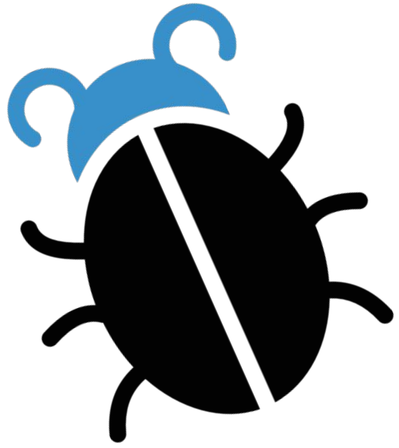
1. Data Acquisition, input or creation
   a. Collecting data from a primary source, such as field observation, experimental research, or surveys
   b. Acquiring data from an existing source, through web-scraping or or generating data via simulation
   c. Regardless of the method, the end result of this first stage is raw data
2. Data Processing or Cleaning
   a. Manual data entry, visual data review, systematic data manipulation or filtering using scripts or software
   b. Relevant data should be digitized, cleaned, and fully prepared for analysis
3. Data Analysis
   a. Formal statistics, data visualization, assessing the performance of particular algorithms, extending the data to address a hypothesis or draw a scientific conclusion
   b. Produces the desired scientific products of the research

Project Example: Collecting Tomatoes

A study in which we have collected field data on tomatoes being grown as part of an agricultural experiment comparing conventional (using fertilizers and pesticides) and organic management.



Measurements of the total yield of tomatoes, in kilograms per plant; produced by four plants in each of three fields having no management after planting (N), conventional management (C), or organic management (O); and observance of substantial insect damage was noted on the plant leaves at the time of harvest.



Data Acquisition -- Raw Data

Field	Weight	Insect
N	5.8	Y
N	5.9	N
N	1.6	Y
N	4.0	Y
N	2.9	Y
C	12.4	N
C	11.5	N
C	9.3	N
C	NA	N
C	12.1	N
O	9.9	N
O	6.7	N
O	10.6	Y
O	3.7	Y
O	NA	N

README.txt

Data collected by undergraduate assistants to Prof John Smith at the Concord Field Station. All plants were located in Field 3 and chosen for measurement when approximately 12" tall. Yields were recorded in August 2018.

Field codes indicate no treatment (N), conventional (C), and organic (O). Yield is in kg, with NA indicating a plant that died prior to yield measurement. Insect damage assessed visually, Y indicates more than 25% loss of leaf area.

```
|-- tomato_project
|    |-- data_raw
|    |    |-- raw_yield_data.csv
|    |    |-- README.txt
|    |-- data_clean
|    |-- results
 |    |-- src
```

Data Processing -- Cleaning up the data

For this tomato yield data, we can readily write a short script that will read the raw table, remove the rows with `NA` yields and those with a field code of `N`, and save the resulting processed data.

```
### Read in the raw data, assuming we are working in the src directory
raw_yield_data <- read.csv("../data_raw/raw_yield_data.csv")

### Clean the data by removing rows with NA and where 'Field' == N
clean_yield_data <- na.omit(raw_yield_data[raw_yield_data$Field != "N", ])

### Write the clean data to disk
 write.csv(clean_yield_data, "../data_clean/clean_yield_data.csv")
```

Data Processing -- Cleaning up the data

Save this as a script `clean_data.R` in the `src` subfolder.

This will read the table `raw_yield_data.csv` from the `data_raw` subfolder, clean it, and save the resulting cleaned table as `clean_yield_data.csv` in the `data_clean` subfolder.

The cleaned data are placed in a different subfolder from the raw data to ensure that the original, raw data are never confused with any derived data products.

Ideally the raw data files should never be altered, with all changes and modifications saved to a separate file.

```
|-- tomato_project
|   |-- data_raw
|   |   |-- raw_yield_data.csv
|   |   |-- README.txt
|   |-- data_clean
|   |   |-- clean_yield_data.csv
|   |-- results
|   |-- src
|   |   |-- clean_data.R
```

Data Processing -- OpenRefine

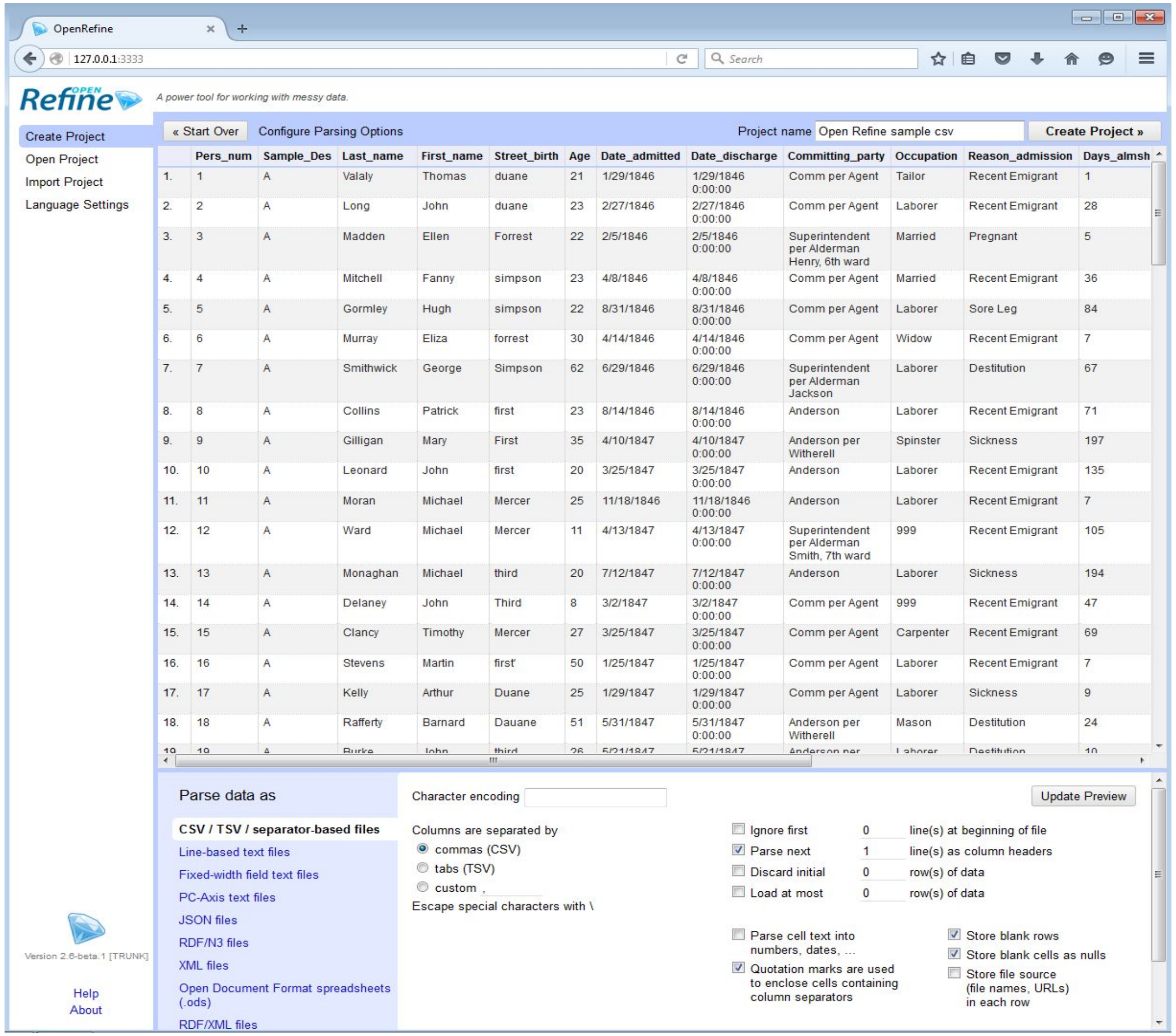
OpenRefine, formerly Google Refine, is an open source tool that allows users to load data, clean it quickly and accurately, transform it, and even geocode it.

- Simple, easy installation
- Lots of great import formats: .tsv, .csv, XML, RDF Triples, JSON, Google Sheets, Excel
- Upload from local drive or import from URL
- Many export formats: .tsv, .csv, Excel, HTML table
- Useful extensions: geoXtension, Opentree for phylogenetic trees from Open Tree of Life

Data Processing -- OpenRefine

A local server will be launched in a terminal and then a browser window will open in your default browser to begin session. (Note: if a window does not open, open a new browser window and visit the URL http://127.0.0.1:3333/)

You preview the data to make sure the import is correct before going into the space where you can clean, munge, and wrangle your data!



Data Processing -- OpenRefine

Some popular uses:

- De-duplicating data
- Geocoding data
- Recode data (changing NULLs to 999, for instance)
- Standardizing data
- Setting data types in bulk (re-encoding data)
- Advanced operations with Regex, Jython, or General Refine Expression Language

Data Analysis -- Working with the data

This next script will read the cleaned data table, perform the desired t-test, and save the summarized results of the test in the `results` subfolder as a plain text file `test_results.txt`.

```
### Load clean data, assuming we are in the src directory
clean_yield_data <- read.csv("../data_clean/clean_yield_data.csv")


### t-test of Weights by Field type: is there significant difference in
### tomato yield in the different fields?
t_test_Weight_Field <- with(clean_yield_data, t.test(Weight ~ Field)


### Write test result to plain text file
capture.output(t_test_Weight_Field, file = "../results/test_results.txt")
```

Data Analysis -- Working with the data

The `test_results.txt` file indicates that there is no detectable significant difference between the yields in the conventional and organic fields (p = 0.104).

Save this script titled `analysis.R` in the `src` directory.

While the code itself is designed to reproduce the quantitative results of an analysis, ***code comments*** and ***other documentation*** are designed to help another researcher reproduce the thought process that went into structuring and writing code in a particular way.

```
|-- tomato_project
|    |-- data_raw
|    |    |-- raw_yield_data.csv
|    |    |-- README.txt
|    |-- data_clean
|    |    |-- clean_yield_data.csv
|    |-- results
|    |    |-- test_results.txt
|    |-- src
|    |    |-- analysis.R
|    |    |-- clean_data.R
```

Data Analysis -- Automation

Create a script that can execute, in one step, all of the various subcomponents of the entire workflow.

In this simple example, our workflow has only two steps that can be performed automatically: executing `clean_data.R` to generate the cleaned data table, and then executing `analysis.R` to perform the statistical test.

Create this shell script, `runall.sh` saved in the `src` directory.

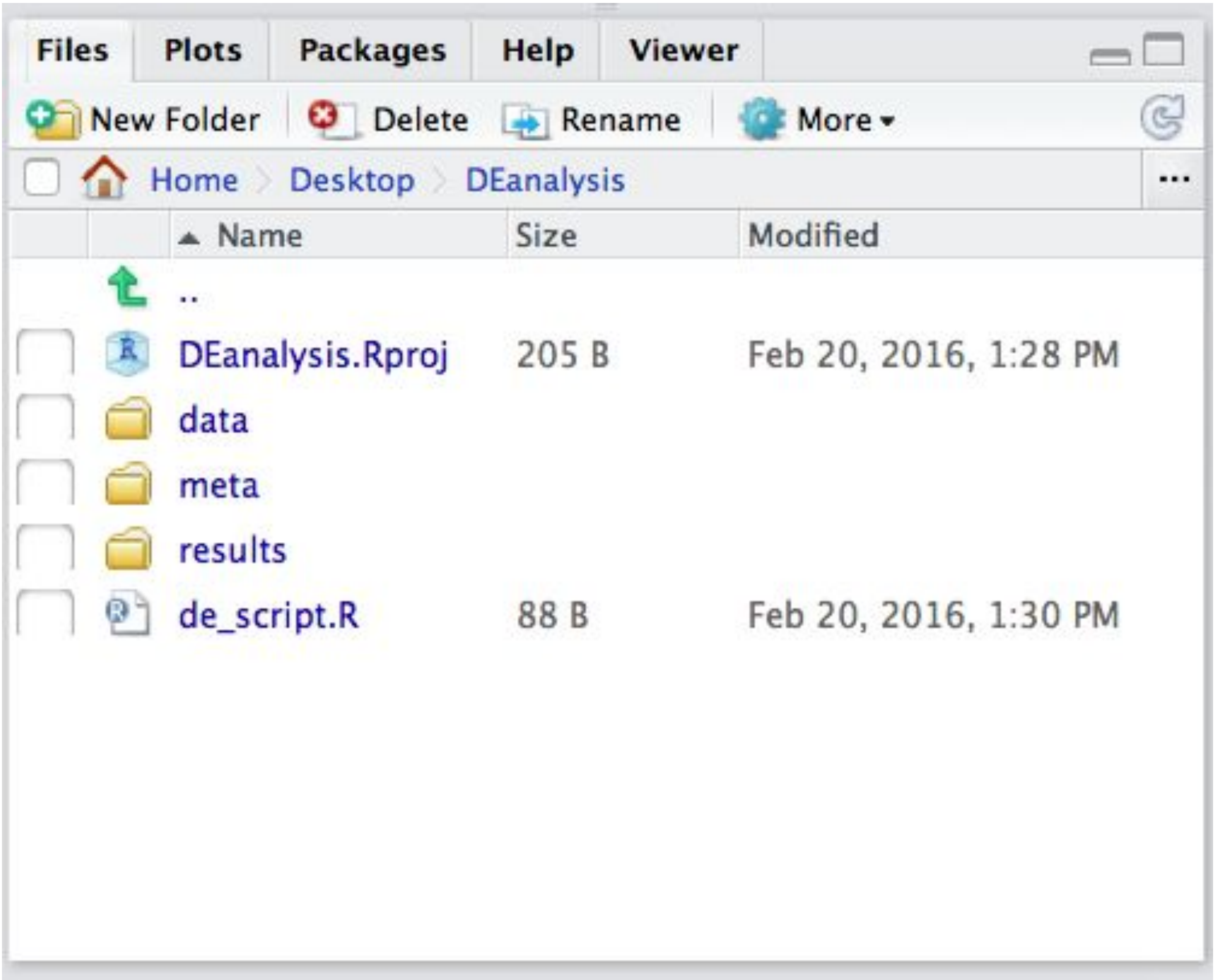
```
r clean_data.R

r analysis.R
```

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

Summary of Digital Gene Expression Analysis Workflow

Name	Length	EffectiveLength	TPM	NumReads
ENST00000456328	1657	1785.304	0.054490	3.722479
ENST00000450305	632	250.000	0.000000	0.000000
ENST00000488147	1351	1530.937	3.793490	222.229533
ENST00000619216	68	3.000	34.844416	4.000000
ENST00000473358	712	519.262	0.000000	0.000000
ENST00000469289	535	250.000	0.000000	0.000000
ENST00000607096	138	5.000	0.000000	0.000000
ENST00000417324	1187	250.000	0.000000	0.000000
ENST00000461467	590	250.000	0.000000	0.000000
ENST00000606857	840	250.000	0.000000	0.000000
ENST00000642116	1414	250.000	0.000000	0.000000
ENST00000492842	939	250.000	0.000000	0.000000



What to capture in a README: The effective gene length in a sample is then the average of the transcript lengths after weighting for their relative expression. The pseudocounts generated by Salmon are represented as normalized TPM (transcripts per million) counts and map to transcripts.

Summary of Digital Gene Expression Analysis Workflow

Summary of differential expression analysis workflow

We have detailed the various steps in a differential expression analysis workflow, providing theory with example code. To provide a more succinct reference for the code needed to run a DGE analysis, we have summarized the steps in an analysis below:

0. Obtaining gene-level counts from Salmon using tximport

```
# Run tximport
txi <- tximport(files, type="salmon", tx2gene=t2g, countsFromAbundance = "lengthScaledTPM")

# "files" is a vector wherein each element is the path to the salmon quant.sf file, and each element is named
# "t2g" is a 2 column data frame which contains transcript IDs mapped to geneIDs (in that order)
```

1. Creating the dds object:

```
# Check that the row names of the metadata equal the column names of the **raw counts** data
all(colnames(txi$counts) == rownames(metadata))

# Create DESeq2Dataset object
dds <- DESeqDataSetFromTximport(txi, colData = metadata, design = ~ condition)
```

2. Exploratory data analysis (PCA & heirarchical clustering) - identifying outliers and sources of variation in the data:

```
# Transform counts for data visualization
rld <- rlog(dds, blind=TRUE)

# Plot PCA
plotPCA(rld, intgroup="condition")

# Extract the rlog matrix from the object and compute pairwise correlation values
rld_mat <- assay(rld)
rld_cor <- cor(rld_mat)

# Plot heatmap
pheatmap(rld_cor, annotation = metadata)
```

3. Run DESeq2:

```
    # **Optional step** - Re-create DESeq2 dataset if the design formula has changed after QC analysis in

# Run DESeq2 differential expression analysis
dds <- DESeq(dds)

    # **Optional step** - Output normalized counts to save as a file to access outside RStudio using "norm
```

4. Check the fit of the dispersion estimates:

```
# Plot dispersion estimates
plotDispEsts(dds)
```

5. Create contrasts to perform Wald testing on the shrunken log2 foldchanges between specific conditions:

```
# Output results of Wald test for contrast
contrast <- c("condition", "level_to_compare", "base_level")
res <- results(dds, contrast = contrast, alpha = 0.05)
res <- lfcShrink(dds, contrast = contrast, res=res)
```

6. Output significant results:

```
# Set thresholds
padj.cutoff < - 0.05

# Turn the results object into a tibble for use with tidyverse functions
res_tbl <- res %>%
              data.frame() %>%
              rownames_to_column(var="gene") %>%
              as_tibble()

# Subset the significant results
sig_res <- filter(res_tbl, padj < padj.cutoff)
```

7. Visualize results: volcano plots, heatmaps, normalized counts plots of top genes, etc.

8. Perform analysis to extract functional significance of results: GO or KEGG enrichment, GSEA, etc.

9. Make sure to output the versions of all tools used in the DE analysis:

```
sessionInfo()
```

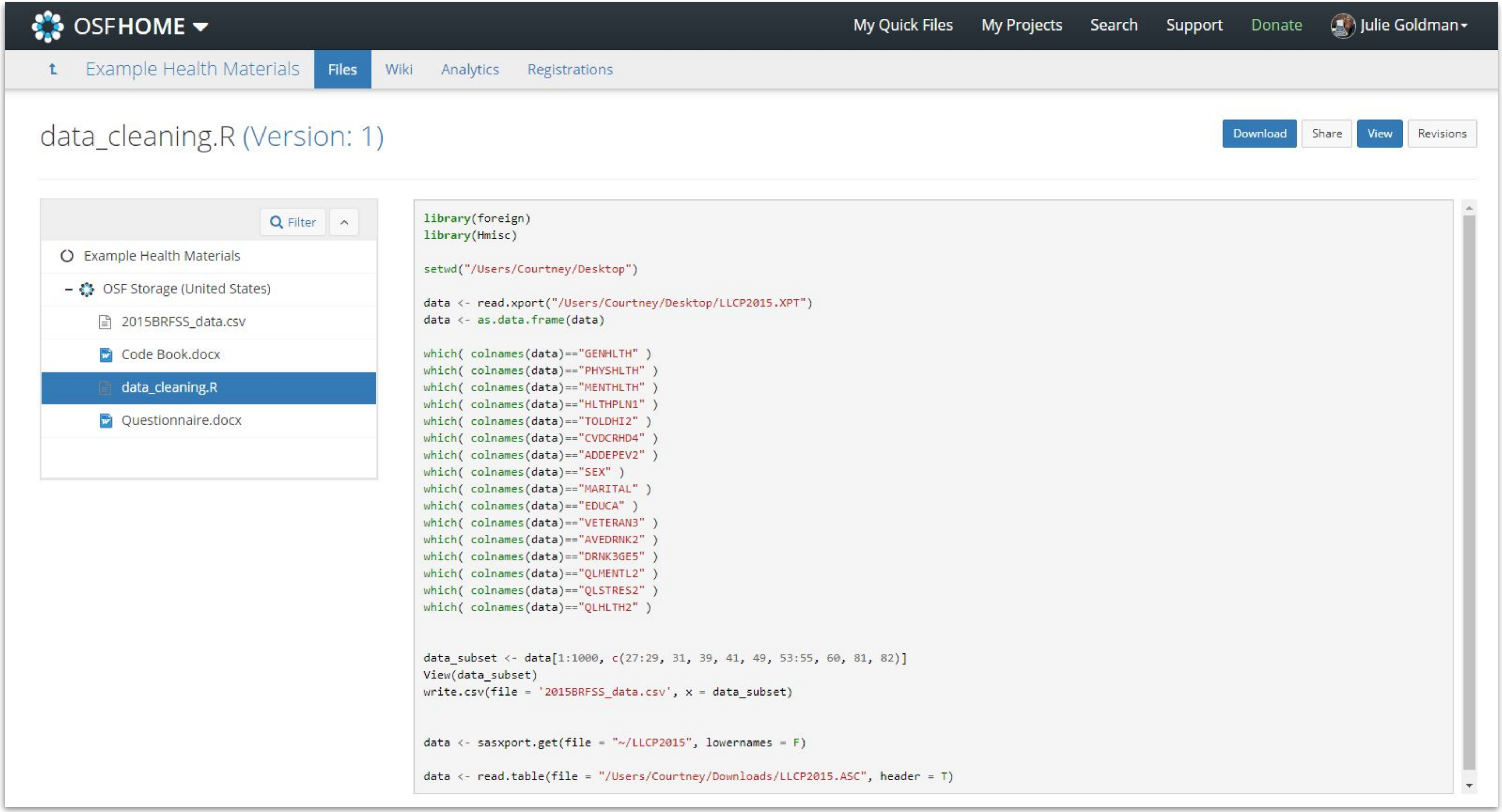
Toolkit for Reproducibility

- **Version control & File management**: *Git*, GitHub, Bitbucket
- **Organizing Your Lab**: Electronic Lab Notebook, *Open Science Framework*
- **Coding**: *R*, RStudio, Python, IPython, Binder
- **Compiling**: GNU Make, Pandoc, ReproZip
- **Writing**: Overleaf, R Markdown, LaTeX, *Jupyter Notebook*
- **Sharing**: Data Repositories, *Dataverse*, *protocol.io*, reagent sharing

Open Science Framework

- OSF is a free, open source, online framework for researchers
  - Accommodates any discipline by allowing you to structure projects to suit your needs
- OSF makes it simple to
  - Organize your research
  - Connect your workflow
  - Keep track of changes to your project
  - Share materials and information with colleagues or the public
- Researchers use OSF to
  - Manage individual and collaborative projects
  - Maintain visibility on multiple file types or components
  - Set up templates for their classrooms or labs for consistent structure of research outputs
  - As an ongoing repository for long-term research data collection
  - As a final home for a research output like a preprint or conference talk

Open Science Framework

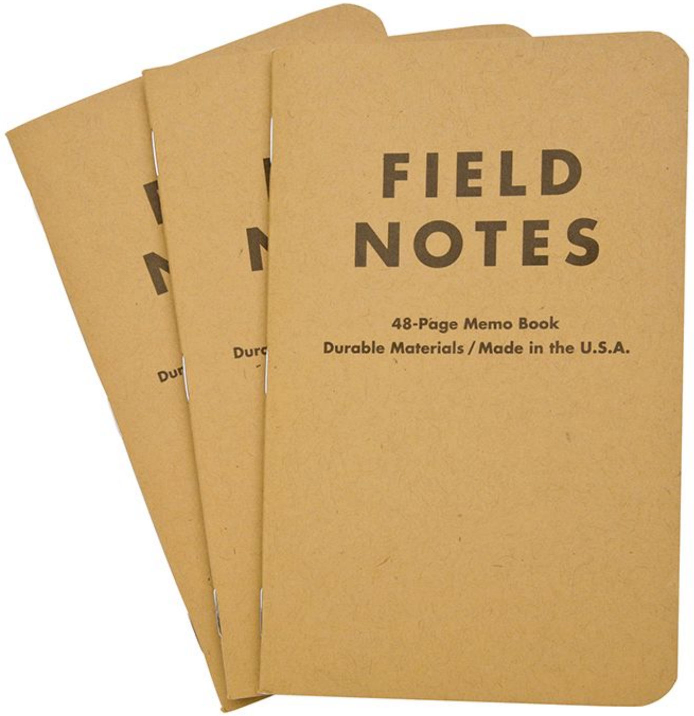


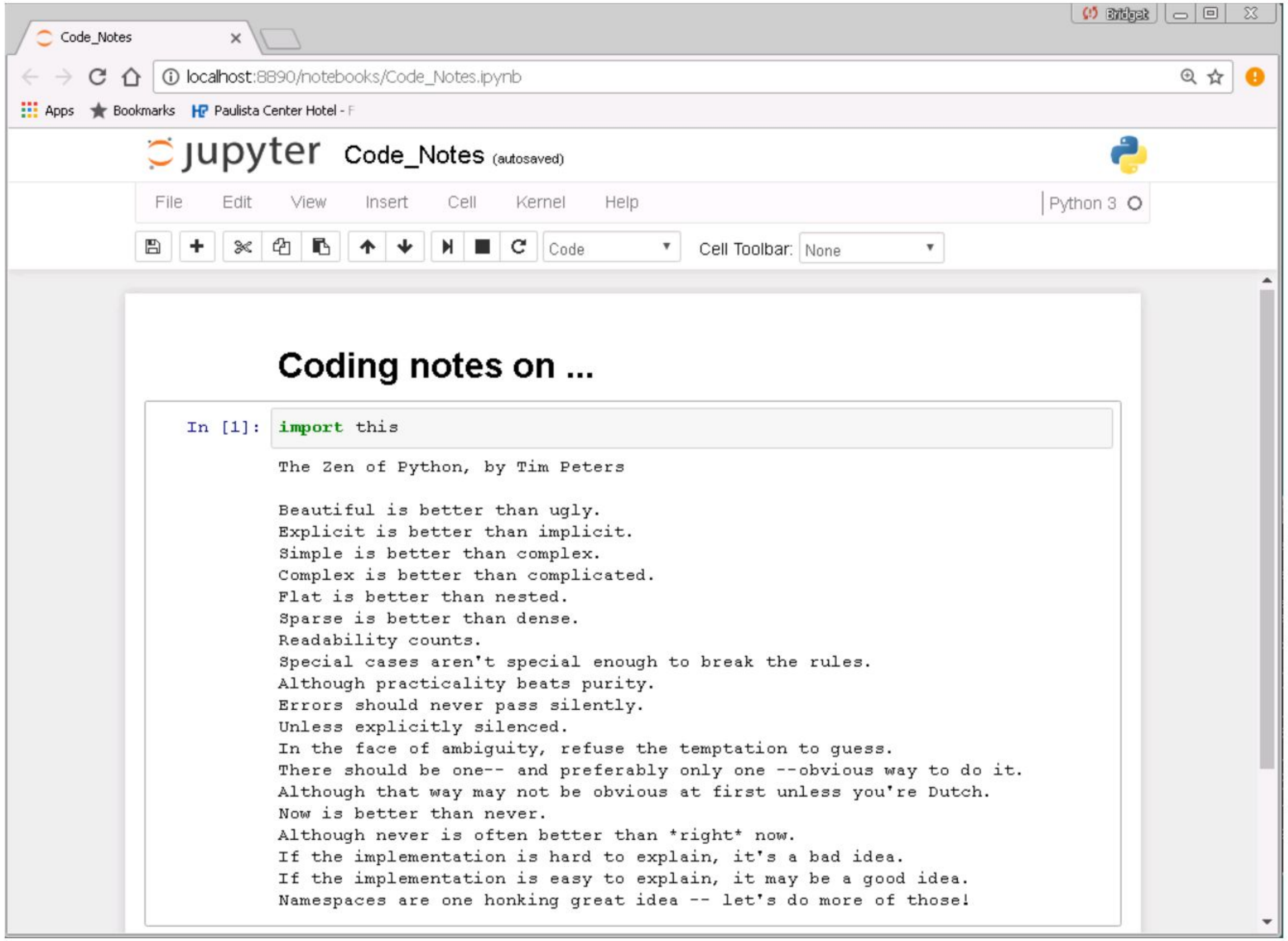
Jupyter Notebook = Code + Documentation

You can think of the notebook as a lab or field notebook that keeps a detailed record of the steps you take as you develop scripts and programming workflows.

Just as you would with a lab notebook, it is important to develop good note-taking habits. It is important to develop the skills, tools, and best practices needed to implement in your own research to enhance reproducibility, which will make modifications, collaboration, and publishing easier.

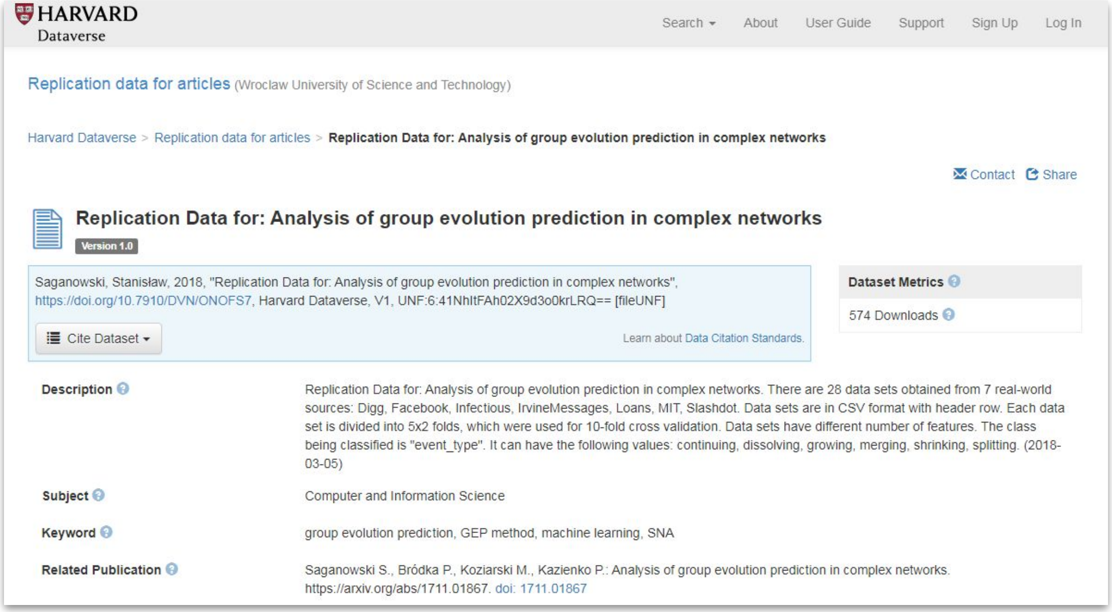
Jupyter Notebook = Code + Documentation





Dataverse Replication Data

- Be sure to include all the necessary descriptive metadata that would make it easier for other researchers to discover your replication dataset.
- When you are ready to upload your replication dataset files into Dataverse, make sure you have:
    - a list of code, scripts, documents and data files
    - include at least a subset(s) of the original dataset files, containing only those variables necessary for replication
    - deposit preferred or commonly-used file formats in your discipline, and remember to remove information from your datasets that must remain confidential
    - sets of computer program recodes
    - program commands, code or script for analysis
    - extracts of existing publicly available data
    - documentation files



Dataverse Replication Data

HARVARD Dataverse

Search ▾ About User Guide Support Sign Up Log In

Replication data for articles (Wroclaw University of Science and Technology)

Harvard Dataverse > Replication data for articles > **Replication Data for: Analysis of group evolution prediction in complex networks**

Contact Share

Replication Data for: Analysis of group evolution prediction in complex networks

Version 1.0

Saganowski, Stanisław, 2018, "Replication Data for: Analysis of group evolution prediction in complex networks", https://doi.org/10.7910/DVN/ONOFS7, Harvard Dataverse, V1, UNF:6:41NhItFAh02X9d3o0krLRQ== [fileUNF]

Cite Dataset ▾

Learn about Data Citation Standards.

Dataset Metrics

574 Downloads

Description	Replication Data for: Analysis of group evolution prediction in complex networks. There are 28 data sets obtained from 7 real-world sources: Digg, Facebook, Infectious, IrvineMessages, Loans, MIT, Slashdot. Data sets are in CSV format with header row. Each data set is divided into 5x2 folds, which were used for 10-fold cross validation. Data sets have different number of features. The class being classified is "event_type". It can have the following values: continuing, dissolving, growing, merging, shrinking, splitting. (2018-03-05)
Subject	Computer and Information Science
Keyword	group evolution prediction, GEP method, machine learning, SNA
Related Publication	Saganowski S., Bródka P., Koziarski M., Kazienko P.: Analysis of group evolution prediction in complex networks. https://arxiv.org/abs/1711.01867. doi: 1711.01867

Methods Sharing

Daniel Gonzales
@dgonzales1990
Folge ich

2017: “Devices were fabricated as previously described [ref 8]”

[ref 8] 2015: “Devices were fabricated as previously described [ref 4]”

[ref 4] 2013: “Devices were fabricated as previously described [ref 2]”

[ref 2] 2009: “Devices were fabricated with conventional methods”

Tweet übersetzen

13:16 - 17. Jan. 2018

230 Retweets 798 „Gefällt mir“-Angaben

28 230 798

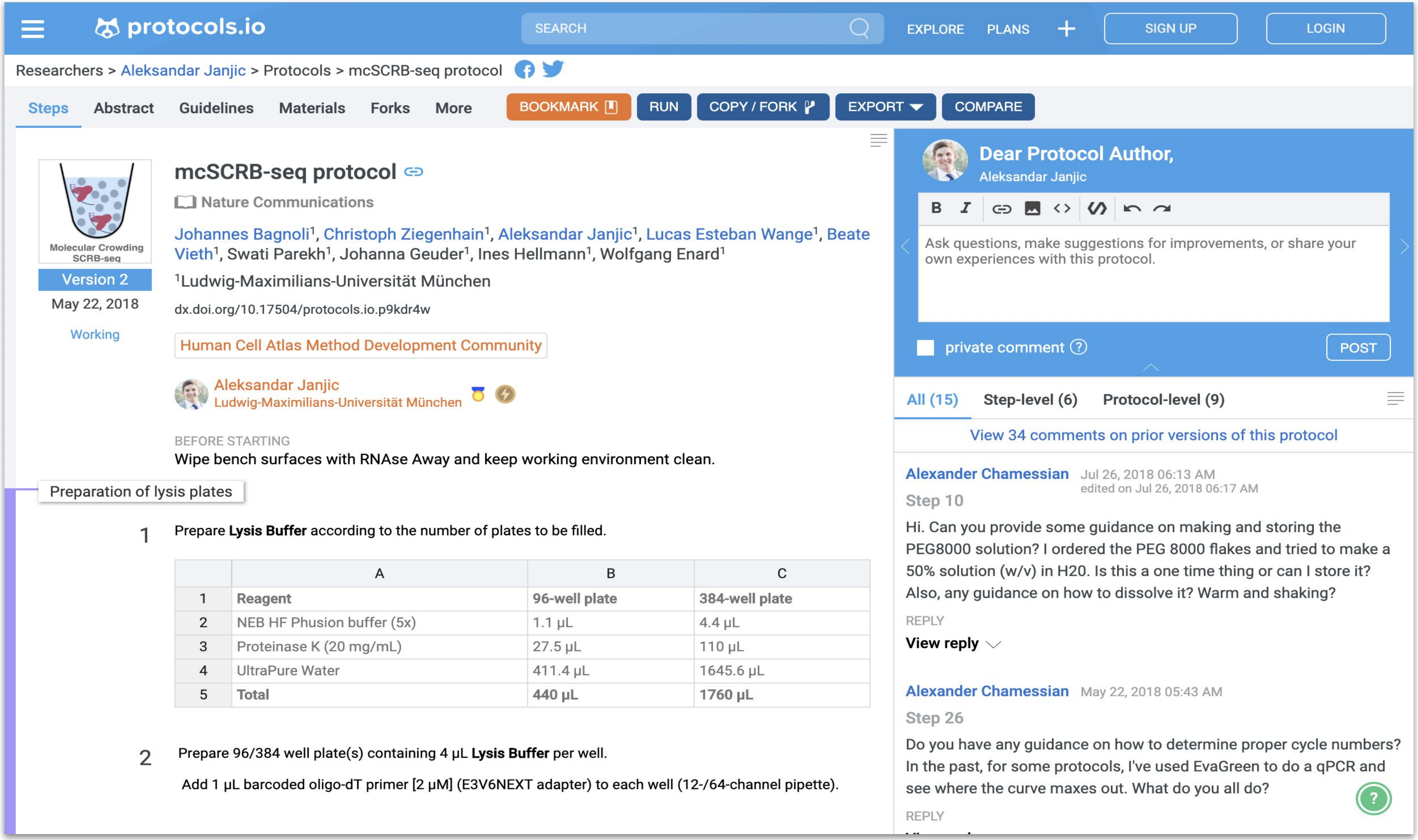
Morgan Halane
@themorgantrail
Follow

Looking for protocol in 1997 paper: "as described in (x) et al '96". Finds '96 paper: "as described in (x) '87." Finds '87 paper: Paywall.

9:20 PM - 1 Nov 2017 from Pohang-si, Republic of Korea

35 Retweets 83 Likes

protocols.io for Methods Sharing



Reproducibility Case Study Example

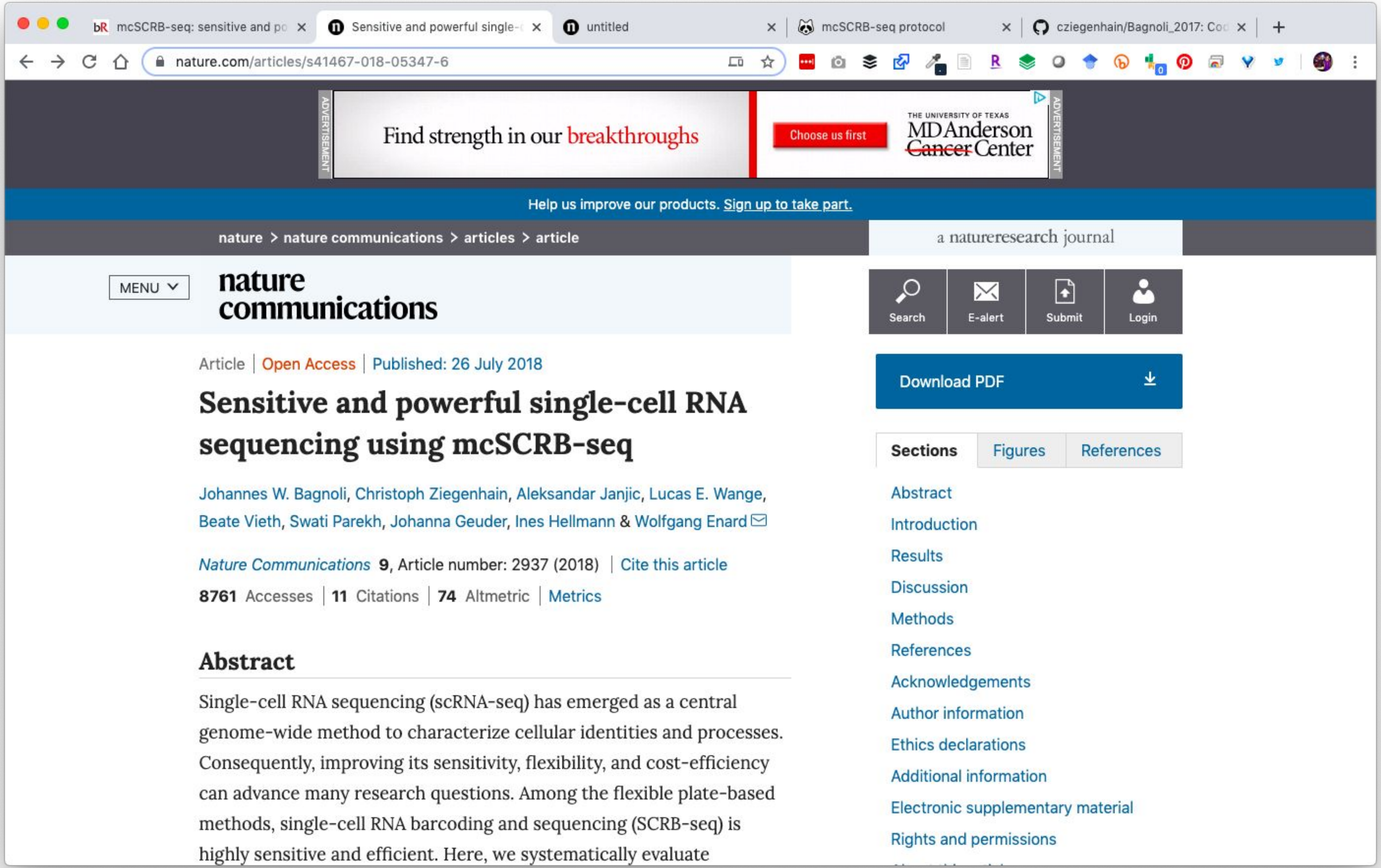
mcSCRB-seq: sensitive and powerful single-cell RNA sequencing

Starting point: https://doi.org/10.1101/188367 (or use bit.ly/repro-rna)

From the preprint, can you locate...

- The published manuscript?
- The code?
- The data?
- The protocol?


bR mcSCRB-seq: sensitive and po
Sensitive and powerful single-
mcSCRB-seq protocol
cziegenhain/Bagnoli_2017: Co
biorxiv.org/content/10.1101/188367v1
Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS
Search
Advanced Search
New Results
Comment on this paper
Previous
Next
mcSCRB-seq: sensitive and powerful single-cell RNA sequencing
Posted October 18, 2017.
Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Lucas E. Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann, Wolfgang Enard
doi: https://doi.org/10.1101/188367
Now published in Nature Communications doi: 10.1038/s41467-018-05347-6
Download PDF
Supplementary Material
Email
Share
Citation Tools
Tweet
Like 0
Abstract
Full Text
Info/History
Metrics
Preview PDF
Subject Area
Genomics
Summary
Single-cell RNA sequencing (scRNA-seq) has emerged as the central genome-wide method to characterize cellular identities and processes. While performance of scRNA-seq methods is improving, an optimum in terms of sensitivity, cost-efficiency and flexibility has not yet been reached. Among the flexible plate-based methods “Single-Cell RNA-Barcoding and Sequencing” (SCRB-seq) is one of the most sensitive and efficient ones. Based on this protocol, we systematically evaluated experimental conditions such as reverse transcriptases, reaction enhancers and PCR polymerases. We find that adding polyethylene glycol considerably increases sensitivity by enhancing cDNA synthesis. Furthermore, using Terra polymerase increases efficiency
Subject Areas
All Articles
Animal Behavior and Cognition
Biochemistry
Bioengineering
Bioinformatics
Biophysics



Help us improve our products. Sign up to take part.

nature > nature communications > articles > article

a natureresearch journal

MENU

nature communications

Search E-alert Submit Login

Article | Open Access | Published: 26 July 2018

Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq

Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Lucas E. Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann & Wolfgang Enard

Nature Communications **9**, Article number: 2937 (2018) | Cite this article

Download PDF

Abstract

Single-cell RNA sequencing (scRNA-seq) has emerged as a central genome-wide method to characterize cellular identities and processes. Consequently, improving its sensitivity, flexibility, and cost-efficiency can advance many research questions. Among the flexible plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq) is highly sensitive and efficient. Here, we systematically evaluate

Batch effect analysis

In order to detect genes differing between batches of one scRNA-seq protocol, data were normalized using scran[31]. Next, we tested for differentially expressed genes using limma-voom[33,34]. Genes were labeled as significantly differentially expressed between batches with Benjamini–Hochberg adjusted *p* values <0.01.

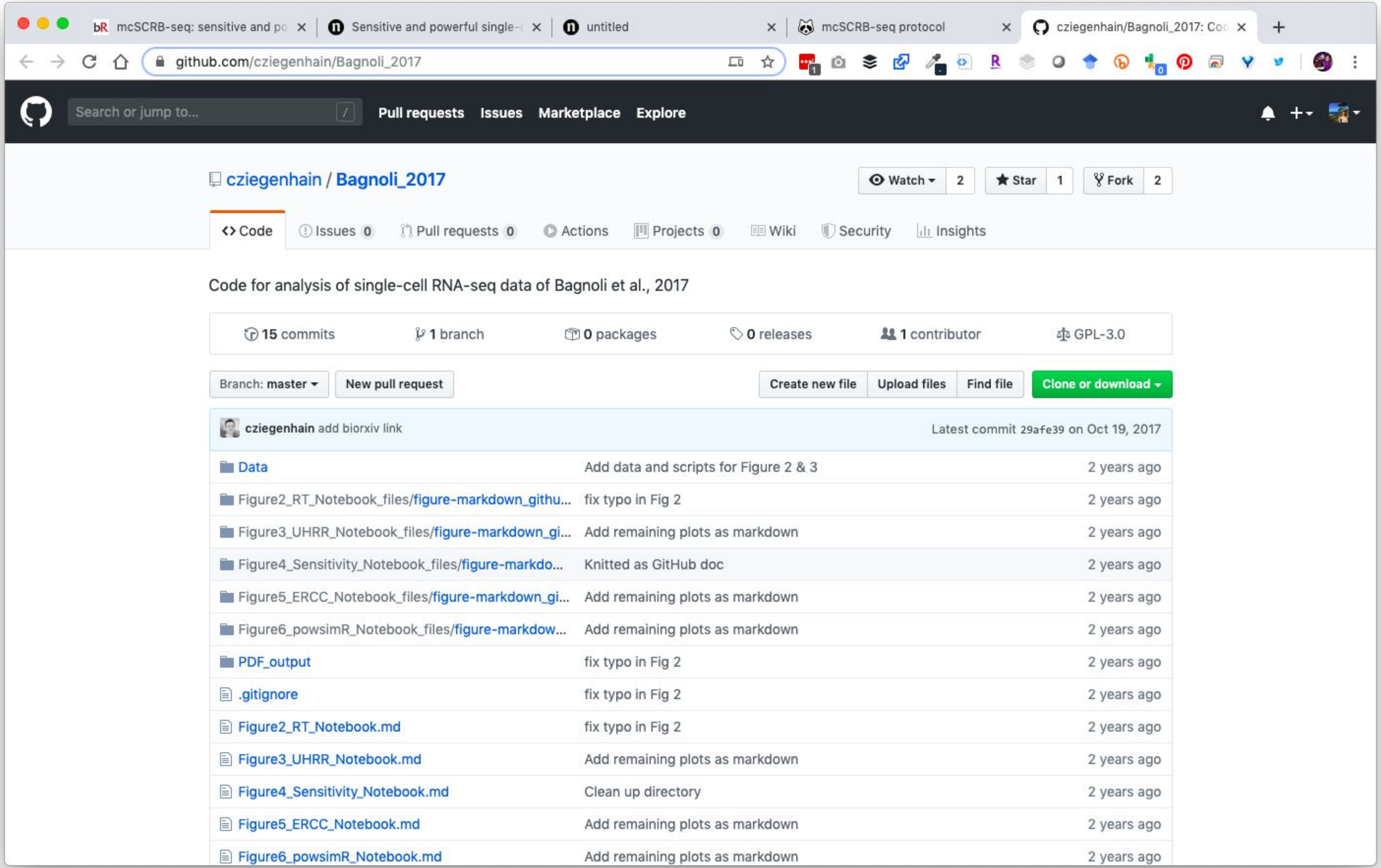
Code availability

Analysis code to reproduce major analyses can be found at https://github.com/cziegenhain/Bagnoli_2017.

Data availability

RNA-seq data generated here are available at GEO under accession GSE103568.

Further data including cDNA yield of optimization experiments is available on GitHub (https://github.com/cziegenhain/Bagnoli_2017). A detailed step-by-step protocol for mcSCRB-seq has been submitted to the protocols.io repository (mcSCRB-seq protocol 2018). All other data available from the authors upon reasonable request.


mcSCRB-seq: sensitive and po
Sensitive and powerful single-
untitled
mcSCRB-seq protocol
cziegenhain/Bagnoli_2017: Co
github.com/cziegenhain/Bagnoli_2017
Search or jump to...
Pull requests
Issues
Marketplace
Explore
cziegenhain / Bagnoli_2017
Watch 2
Star 1
Fork 2
Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Code for analysis of single-cell RNA-seq data of Bagnoli et al., 2017
15 commits
1 branch
0 packages
0 releases
1 contributor
GPL-3.0
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
cziegenhain add biorxiv link
Latest commit 29afe39 on Oct 19, 2017
Data | Add data and scripts for Figure 2 & 3 | 2 years ago
Figure2_RT_Notebook_files/figure-markdown_githu... | fix typo in Fig 2 | 2 years ago
Figure3_UHRR_Notebook_files/figure-markdown_gi... | Add remaining plots as markdown | 2 years ago
Figure4_Sensitivity_Notebook_files/figure-markdo... | Knitted as GitHub doc | 2 years ago
Figure5_ERCC_Notebook_files/figure-markdown_gi... | Add remaining plots as markdown | 2 years ago
Figure6_powsimR_Notebook_files/figure-markdow... | Add remaining plots as markdown | 2 years ago
PDF_output | fix typo in Fig 2 | 2 years ago
.gitignore | fix typo in Fig 2 | 2 years ago
Figure2_RT_Notebook.md | fix typo in Fig 2 | 2 years ago
Figure3_UHRR_Notebook.md | Add remaining plots as markdown | 2 years ago
Figure4_Sensitivity_Notebook.md | Clean up directory | 2 years ago
Figure5_ERCC_Notebook.md | Add remaining plots as markdown | 2 years ago
Figure6_powsimR_Notebook.md | Add remaining plots as markdown | 2 years ago

mcSCRB-seq: sensitive a | mcSCRB-seq: sensitive a | Sensitive and powerful s | GEO Accession viewer | mcSCRB-seq protocol | cziegenhain/Bagnoli_201

ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103568&holding=hulib

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

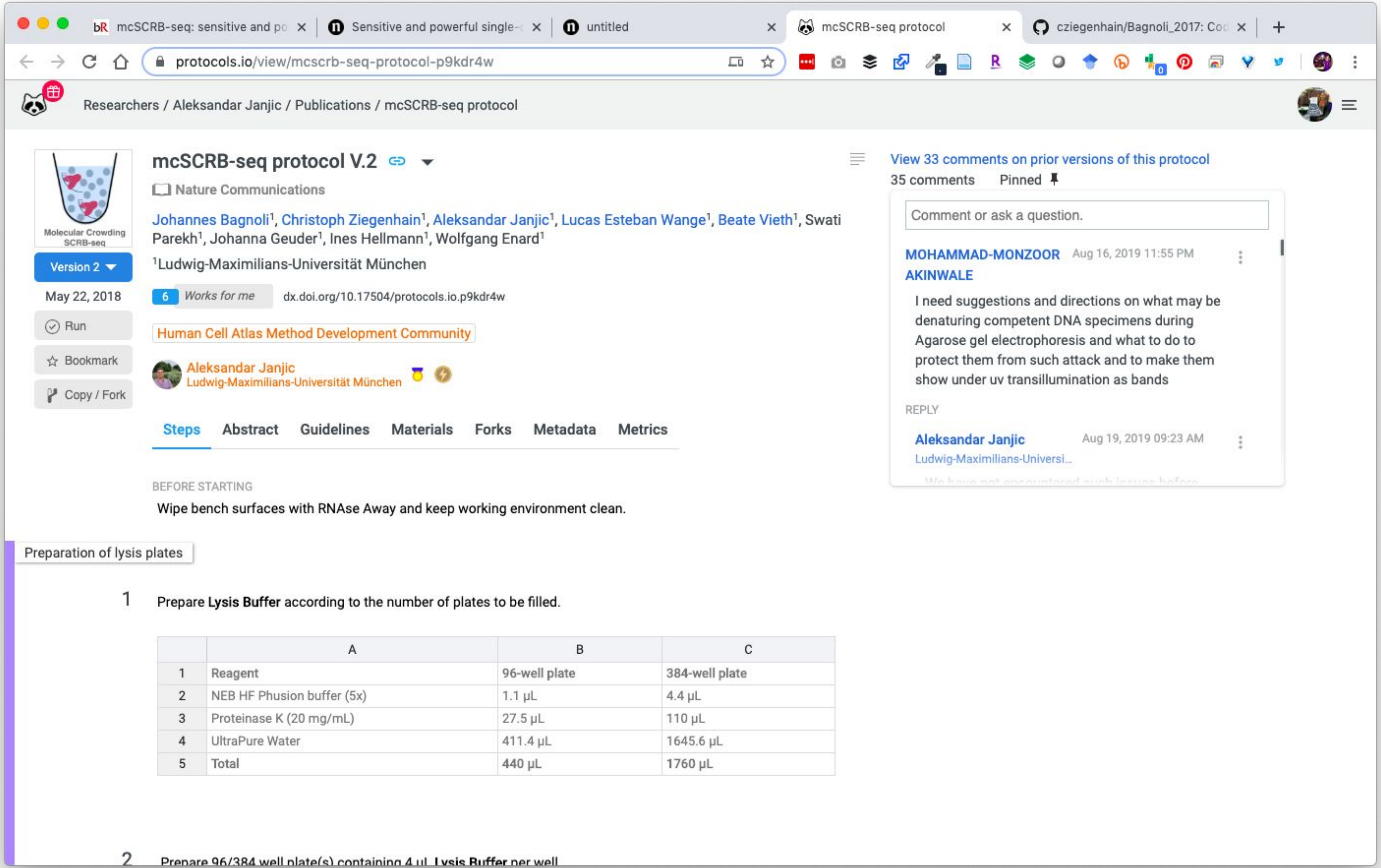
NCBI > GEO > **Accession Display** | Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE103568 | GO

Series GSE103568 | Query DataSets for GSE103568

Status	Public on Jul 26, 2018
Title	Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq
Organisms	Homo sapiens; Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Single-cell RNA sequencing (scRNA-seq) has emerged as a central genome-wide method to characterize cellular identities and processes. Consequently, improving its sensitivity, flexibility, and cost-efficiency can advance many research questions. Among the flexible plate-based methods, single-cell RNA barcoding and sequencing (SCRB-seq) is highly sensitive and efficient. Here, we systematically evaluate experimental conditions of this protocol and find that adding polyethylene glycol considerably increases sensitivity by enhancing cDNA synthesis. Furthermore, using Terra polymerase increases efficiency due to a more even cDNA amplification that requires less sequencing of libraries. We combined these and other improvements to develop a scRNA-seq library protocol we call molecular crowding SCRB-seq (mcSCRB-seq), which we show to be one of the most sensitive, efficient, and flexible scRNA-seq methods to date.
Overall design	single-cell RNA-sequencing protocols were benchmarked on J1, JM8 and Universal Human Reference RNA (UHRR) and human PBMCs
Contributor(s)	Ziegenhain C, Bagnoli JW
Citation(s)	Bagnoli JW, Ziegenhain C, Janjic A, Wange LE et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat Commun* 2018 Jul 26;9(1):2937. PMID: 30050112
Submission date	Sep 06, 2017
Last update date	May 15, 2019
Contact name	Christoph Ziegenhain
E-mail(s)	ziegenhain@bio.lmu.de
Organization name	Ludwig-Maximilians University Munich
Department	Department Biology II
Lab	Anthropology & Human genomics


mcSCRB-seq: sensitive and po
Sensitive and powerful single-
untitled
mcSCRB-seq protocol
cziegenhain/Bagnoli_2017: Co
protocols.io/view/mcscrb-seq-protocol-p9kdr4w
Researchers / Aleksandar Janjic / Publications / mcSCRB-seq protocol
Molecular Crowding SCRB-seq
Version 2
May 22, 2018
Run
Bookmark
Copy / Fork
mcSCRB-seq protocol V.2
Nature Communications
Johannes Bagnoli[1], Christoph Ziegenhain[1], Aleksandar Janjic[1], Lucas Esteban Wange[1], Beate Vieth[1], Swati Parekh[1], Johanna Geuder[1], Ines Hellmann[1], Wolfgang Enard[1]
[1]Ludwig-Maximilians-Universität München
6 Works for me
dx.doi.org/10.17504/protocols.io.p9kdr4w
Human Cell Atlas Method Development Community
Aleksandar Janjic
Ludwig-Maximilians-Universität München
Steps
Abstract
Guidelines
Materials
Forks
Metadata
Metrics
BEFORE STARTING
Wipe bench surfaces with RNAse Away and keep working environment clean.
Preparation of lysis plates
1 Prepare Lysis Buffer according to the number of plates to be filled.
A B C
1 Reagent 96-well plate 384-well plate
2 NEB HF Phusion buffer (5x) 1.1 µL 4.4 µL
3 Proteinase K (20 mg/mL) 27.5 µL 110 µL
4 UltraPure Water 411.4 µL 1645.6 µL
5 Total 440 µL 1760 µL
View 33 comments on prior versions of this protocol
35 comments
Pinned
Comment or ask a question.
MOHAMMAD-MONZOOR AKINWALE
Aug 16, 2019 11:55 PM
I need suggestions and directions on what may be denaturing competent DNA specimens during Agarose gel electrophoresis and what to do to protect them from such attack and to make them show under uv transillumination as bands
REPLY
Aleksandar Janjic
Aug 19, 2019 09:23 AM
Ludwig-Maximilians-Universi...

The goal is to encourage the community to consciously choose open tools to increase interoperability & sustainability of their research.

Open ≠ Reproducible

- Was the research published?
- Were methods adequately reported?
- Was the analysis plan transparent?
- Was the code, data and materials transparent?
- Is the published article accessible?
- Is the published article discoverable?
- Do incentives align?

Research article

Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*

Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, Gustav Nilsonne, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J. Yoon, Michael Henry Tessler, Richie L. Lenne, Sara Altman, Bria Long and Michael C. Frank

Published: **15 August 2018** | https://doi.org/10.1098/rsos.180448

Abstract

Access to data is a critical feature of an efficient, progressive and ultimately self-correcting scientific ecosystem. But the extent to which in-principle benefits of data sharing are realized in practice is unclear. Crucially, it is largely unknown whether published findings can be reproduced by repeating reported analyses upon shared data ('analytic reproducibility'). To investigate this, we conducted an observational evaluation of a mandatory open data policy introduced at the journal *Cognition*. Interrupted time-series analyses indicated a substantial post-policy increase in data available statements (104/417, 25% pre-policy to 136/174, 78% post-policy), although not all data appeared reusable (23/104, 22% pre-policy to 85/136, 62%, post-policy). For 35 of the articles determined to have reusable data, we attempted to reproduce 1324 target values. Ultimately, 64 values could not be reproduced within a 10% margin of error. For 22 articles all target values were reproduced, but 11 of these required author assistance. For 13 articles at least one value could not be reproduced despite author assistance. Importantly, there were no clear indications that original conclusions were seriously impacted. Mandatory open data policies can increase the frequency and quality of data sharing. However, suboptimal data curation, unclear analysis specification and reporting errors can impede analytic reproducibility, undermining the utility of data sharing and the credibility of scientific findings.

Hardwicke, Tom E., Maya B. Mathur, Kyle MacDonald, Gustav Nilsonne, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr et al. 2018. "Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition." *Royal Society open science* 5(8): 180448. https://doi.org/10.1098/rsos.180448

Tips for Reproducibility

- Plan for reproducibility before you start
  - Write a study plan or set up an electronic lab notebook
- Keep track of things
  - Track changes using version control and document everything in a README file
- Report your research transparently
  - Share your protocols and write manuscripts collaboratively
- Archive & share your materials
  - Share and license your research products

The most important tool is the mindset, when starting, that the end product will be reproducible.

– Keith Baggerly

Tools, Resources & Activities

- Different elements associated with each lifecycle component
- Vary by discipline and institution

PLAN	COLLECT, GENERATE & STORE	CLEAN, ANALYZE & VISUALIZE	PUBLISH & SHARE	ARCHIVE & PRESERVE	REUSE
RDM plans DMPTool	Instruments Surveys Licensed data Research computing storage	R, Python OpenRefine SPSS, STATA NVivo Tableau	Data curation Data citations, DOIs Data use agreements (DUAs)	Dataverse Appraise data for retention Preserve data	Dataverse Find data for new project

Services Available

Countway Library Digital Scholarship & Communication

- Data management planning and organization
- Digital scholarship support (open access, data sharing)
- Bioinformatics consultation and training

Harvard Chan Bioinformatics Core

- Next generation sequencing analysis
- Functional analysis
- Grant and manuscript support (data submission to GEO, SRA)
- Bioinformatics training program

Research Computing and Information Technology

- Data analysis and visualization support
- High performance computing and storage services
- Scripting & statistical language training classes

Learning Objectives

- Understand the important impact of creating reproducible research
- Establish a reproducible workflow within the context of an example
- Know services and tools available to support reproducible research

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

– Jon Claerbout

Reproducible research practices enables you to:

- Organize experiments productively
- Accurately analyze results
- Share results with future researchers
- Share techniques
- Share reagents with future researchers
- Accelerate science!

Upcoming Seminars

Best Practices for Keeping a Lab Notebook

Friday, November 22, 2019
1:00pm - 2:00pm
Countway Library, Classroom 403

bit.ly/RDM-Seminars

Data Policy Compliance

Wednesday, Wednesday 18, 2019
1:00pm - 2:00pm
Modell 100A Lecture Hall

bit.ly/RDM-Seminars

http://tinyurl.com/hbc-modules

Resources

Steeves, Vicky. 2018. “Building Services Around Reproducibility & Open Scholarship.” OSF. March 5. https://osf.io/pv6ea/

Sayre, Franklin D, and Amy Riegelman. 2019. “Materials (public).” OSF. February 11. https://osf.io/n8dv2/

Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press. https://www.practicereproducibleresearch.org/

Harvard Biomedical Data Management. https://datamanagement.hms.harvard.edu/

Tips for Reproducibility. Harvard Biomedical Data Management. https://datamanagement.hms.harvard.edu/tips-reproducibility

Harvard Chan Bioinformatics Core. http://bioinformatics.sph.harvard.edu/

Harvard Medical School Research Computing. https://rc.hms.harvard.edu/

Data Carpentry. https://datacarpentry.org/lessons/

Software Carpentry. https://software-carpentry.org/lessons/