Data and Code Reproducibility

Water Systems Group Meeting

12/07/2018

VIRGINIA TECH

Biological Systems Engineering

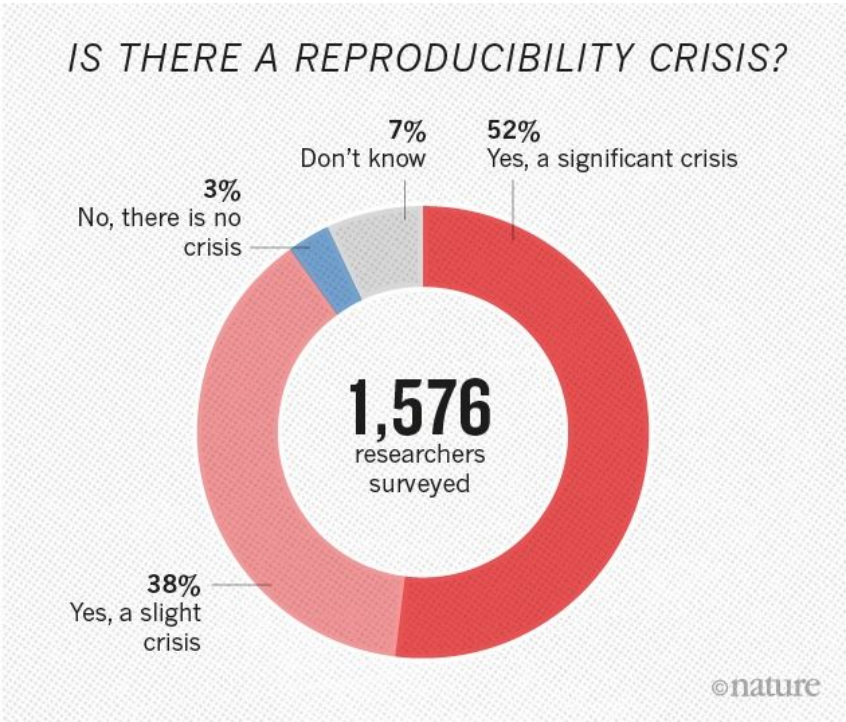
Adapted from April Clyburne-Sherin's Workshop on June 27, 2018

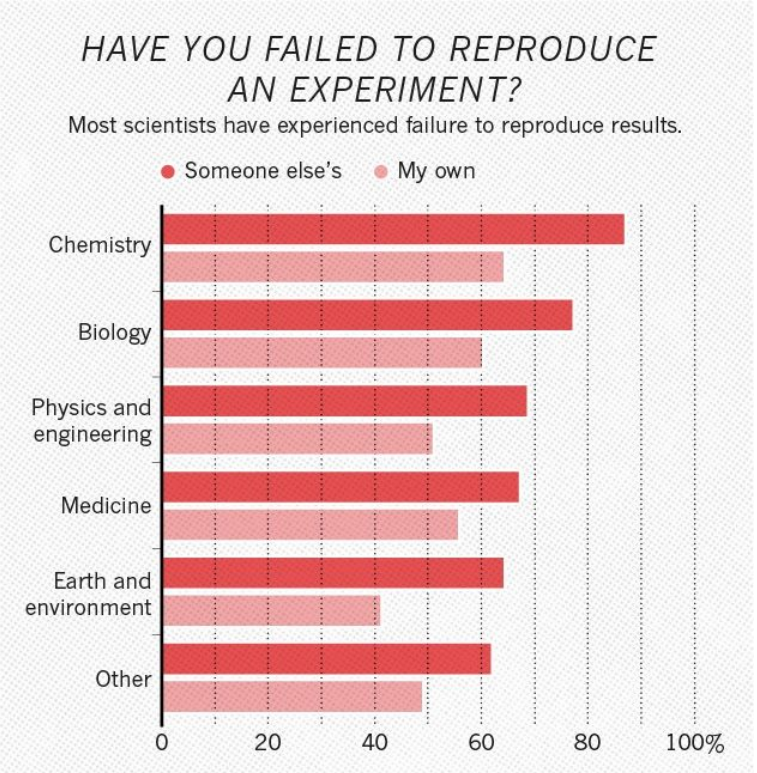
Bad Code


WHY IS THIS STRUCTURE HERE ?
WHERE COULD THIS BRIDGE POSSIBLY LEAD ?
THIS SIGN DOESN'T HELP ME MUCH .
//THIS IS BRIDGE


THE PROJECT I INHERITED HAS WEAK CODE. I NEED TO REWRITE IT FROM SCRATCH.
WILL THERE EVER BE AN ENGINEER WHO SAYS, "THAT LAST GUY DID A GREAT JOB. LET'S KEEP ALL OF IT"?
I'M HOPING THE IDIOT YOU HIRE TO REPLACE ME SAYS THAT.
DilbertCartoonist@gmail.com
Dilbert.com
© 2014 Scott Adams, Inc.

Is This a Problem?

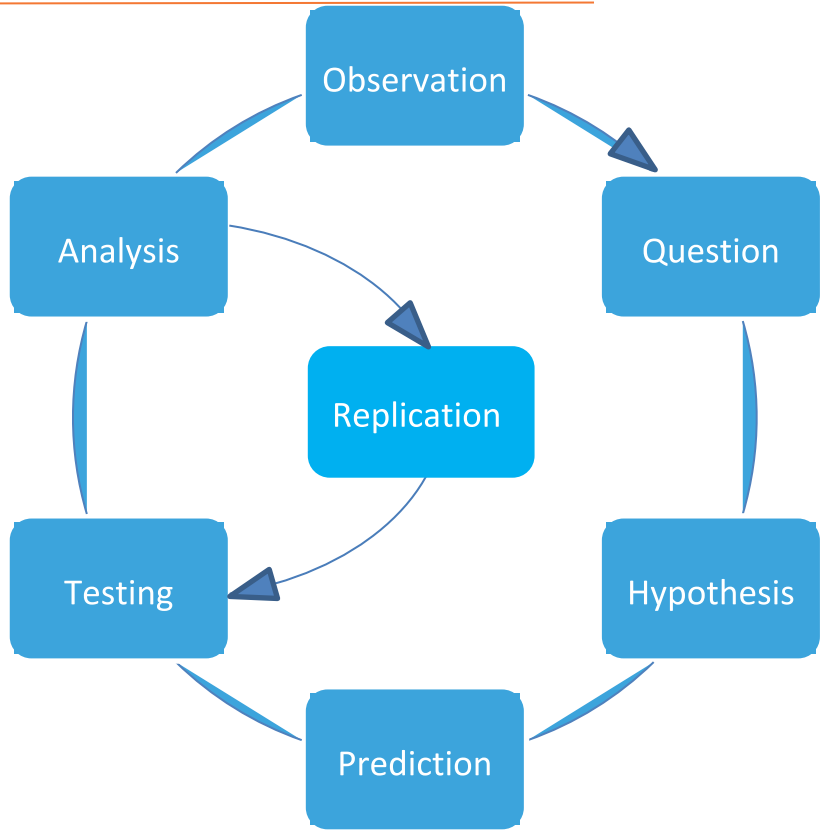




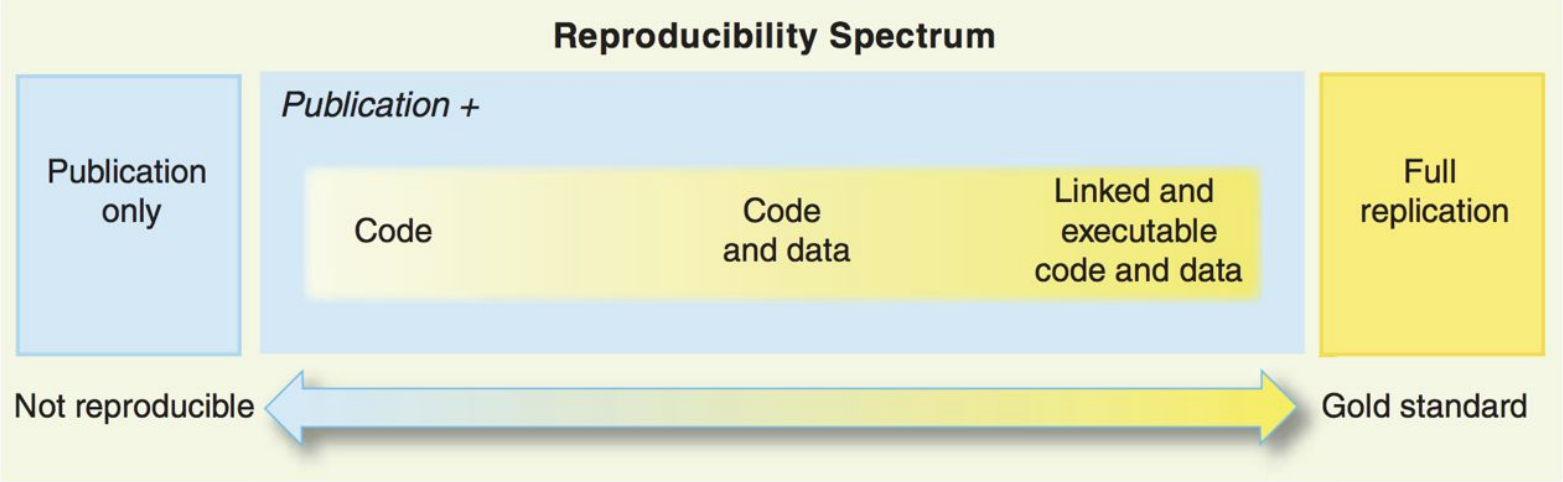
(Nature, 2016)

Importance of Reproducibility

- Results are credible is they are reproducible
- Reusable methods allows for other scientists to build on your research and make new discoveries



Spectrum


Reproducibility Spectrum
Publication only
Publication +
Code
Code and data
Linked and executable code and data
Full replication
Not reproducible
Gold standard

Ten Simple Rules

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Main Themes

Organization

Documentation

Automation

Dissemination

Checklist

```
/ [root]
├── code
│   ├── my_algorithm.py
│   ├── README.md
│   ├── run.sh
│   └── ...
├── data
│   ├── my_data.csv
│   ├── my_sample_image.png
│   └── ...
└── results
    └── [your future results]
```

- Create one repository or directory that holds all related research files.
- Organize your research to separate data, code, and results.
- Save results explicitly.

Tools

OSF

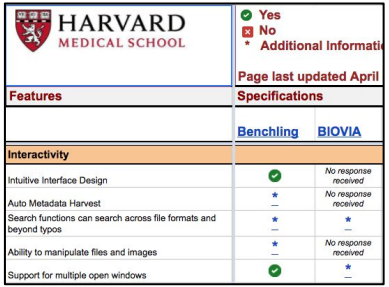
GitHub

CODE OCEAN BETA

SCINOTE

- Open Science Framework: collaborative project organization tool
- GitHub: collaborative coding, and project management
- eLNs: free or paid, lab organization
- Code Ocean: built in best practices

Resources



HARVARD MEDICAL SCHOOL

Yes
No
* Additional Informati

Page last updated April

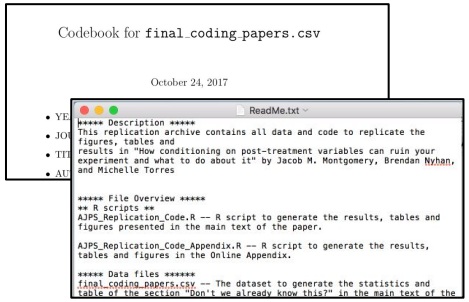
Features	Specifications	
	Benchling	BIOVIA
Interactivity		
Intuitive Interface Design	✓	No response received
Auto Metadata Harvest	*	No response received
Search functions can search across file formats and beyond typos	*	*
Ability to manipulate files and images	*	No response received
Support for multiple open windows	✓	*

- Karl Broman: http://kbroman.org/steps2rr/pages/organize.html
- Harvard eLN Features Matrix: https://docs.google.com/spreadsheets/d/1ar8fgwagOh30E31EAPL-Gorwn_g6XNf81q3VDQnQ_I8/edit?usp=sharing

Have clear sections in code. Always helpful to start with clc and clear

```
%Clear environment
clc
clear
close all
load('Q4data.mat');

%%%% Part 1: euclidean
%load 'neighborhood' function where nearest neighbor criteria is 5 and maximum distance is 6
[p_dist, knn, D]=neighborhood(X_data,6,5);
disp(D(1:8,1:8)); %shows 8x8 Euclidean distance matrix

% 200 points p_dist<6
plot(graph(p_dist==1));

% 200 points k=5
figure;plot(graph(knn==1));%with predefined function this returns matched case


%%%% Part 2: geodesic
[p_dist, knn, D]=neighborhood(X_data(:,1:8),6,2); %k is now changed to 2 to find pair. assuming maximum distance is still 6
geodistance=geodesic(knn,D);
disp(geodistance(1:8,1:8));
```

Checklist



- Document each element or variable in your dataset with a data dictionary / codebook.
- Create a README file.
- Choose licences.
- Consider literate programming.
- Follow FAIR Principles.

Tools

GitHub

jupyter

CODE OCEAN BETA

- Version control: git and GitHub tracks changes to documents and metadata
- Literate programming: knits documentation with code (Jupyter)
- Document & share metadata: Code Ocean renders documentation, notebooks, and records metadata

Resources

DCC because good research needs good data

Popular Licenses

The following OSI-approved licenses are popular, widely used,

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license

- DataONE: https://www.dataone.org/best-practices/create-data-dictionary
- Cornell: https://data.research.cornell.edu/content/readme
- Digital Curation Center: http://www.dcc.ac.uk/resources/how-guides/license-research-data
- OSI: https://opensource.org/licenses

FAIR Principles

TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

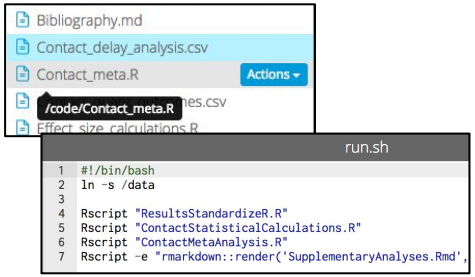
TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

COMMENT COMMENT COMMENT

```
#Create Variables for Desired Information
Facility_Name<-character() #Name of the CWA regulated discharging facility
FacilityID<-character() #Facility ID
VPDESID<-character() #Unique ID used in Virginia for a facility's outfall: concatonated facility ID with 3 digit outfall ID
eff_limit<-numeric() #numerical limit for flow
eff_limit_units<-character() #units of measure applicable to effluent quantity limit
dmr_value<-numeric() #measured effluent through outfall
dmr_units<-character() #units for measured effluent
statistic<-character() #indicates the statistic analysis used for the measured effluent-we are interested in averages
mp_begin<-character() #beginning date of monitoring period (mp)
mp_end<-character() #end data of monitoring period (mp)
mon_in_mp<-numeric() #number of months included in monitoring period
nodi<-character() #if the DMR value is NA, the no data indicator code describes why that is the case
```

Checklist



- Use relative rather than absolute paths.
- Create a master script that runs your scripts in sequence.

Tools

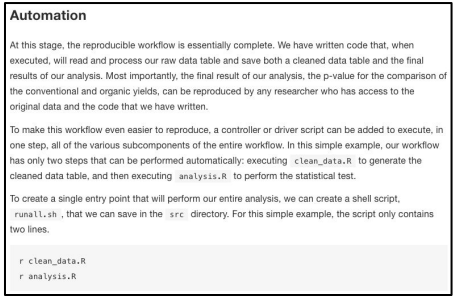
CODE OCEAN

docker

binder (beta)

- Docker: share automated code for devs
- Code Ocean: easy configuring, preservation, & reuse of automated code
- Binder: share automated code for using containers

Resources



Automation

At this stage, the reproducible workflow is essentially complete. We have written code that, when executed, will read and process our raw data table and save both a cleaned data table and the final results of our analysis. Most importantly, the final result of our analysis, the p-value for the comparison of the conventional and organic yields, can be reproduced by any researcher who has access to the original data and the code that we have written.

To make this workflow even easier to reproduce, a controller or driver script can be added to 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.

To create a single entry point that will perform our entire analysis, we can create a shell script, `runall.sh`, that we can save in the `src` directory. For this simple example, the script only contains two lines.

```
r clean_data.R
r analysis.R
```

- Karl Broman on paths: http://kbroman.org/steps2rr/pages/organize.html
- Resource on automation using a master script: https://www.practicereproducibleresearch.org/core-chapters/3-basic.html

```
%Clear environment
clc
clear
close all

%True and False Positive Rates for C1
%TP are on Y-axis while FP are on X-axis
C1_TPR=[1,1,0.8,0.8,0.8,0.6,0.6,0.4,0.2,0.2,0];
C1_FPR=[1,0.8,0.8,0.6,0.4,0.4,0.2,0.2,0.2,0,0];

%True and False Positive Rates for C2
C2_TPR=[1,1,1,1,0.8,0.8,0.8,0.6,0.4,0.2,0];
C2_FPR=[1,0.8,0.6,0.4,0.4,0.2,0,0,0,0,0];

%Baseline
x=[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1];
y=[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1];

%ROC Curves
plot(C1_FPR,C1_TPR,'r')
title('ROC Curves for Models C1 and C2');
xlabel('False Positive Rate');
ylabel('True Positive Rate');
hold on
plot(C2_FPR,C2_TPR,'b')
hold on
plot(x,y,'--k')
legend({'C1','C2','random guess'},'Location','southeast');
hold off

%Area under the curves

Area_C1=trapz(sort(C1_FPR),sort(C1_TPR));
Area_C2=trapz(sort(C2_FPR),sort(C2_TPR));
```

You should be able to press run and be done. Makes your life and everyone else's life easier!

Checklist

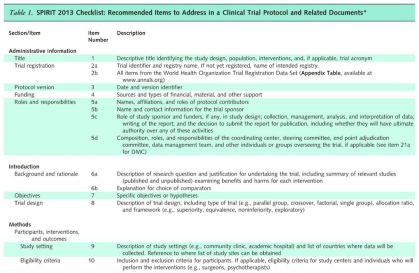


Table 1. SPIRIT 2013 Checklist: Recommended Items to Address in a Clinical Trial Protocol and Related Documents*

Section/Item	Item Number	Description
Administrative information		
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym
Trial registration	2a	Trial identifier and registry name. If not yet registered, name of intended registry
	2b	All items from the World Health Organization Trial Registration Data Set (**Appendix Table**, available at www.annals.org)
Protocol version	3	Date and version identifier
Funding	4	Sources and types of financial, material, and other support
Roles and responsibilities	5a	Names, affiliations, and roles of protocol contributors
	5b	Name and contact information for the trial sponsor
	5c	Role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities
	5d	Composition, roles, and responsibilities of the coordinating center, steering committee, and point adjudication committee, data management team, and other individuals or groups overseeing the trial, if applicable (see item 21a for DMC)
Introduction		
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention
	6b	Explanation for choice of comparators
Objectives	7	Specific objectives or hypotheses
Trial design	8	Description of trial design, including type of trial (e.g., parallel group, crossover, factorial, single group), allocation ratio, and framework (e.g., superiority, equivalence, noninferiority, exploratory)
Methods		
Participants, interventions, and outcomes		
Study setting	9	Description of study settings (e.g., community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centers and individuals who will perform the interventions (e.g., surgeons, psychotherapists)

- Report transparently & completely
  - Write a detailed study protocol before you gather your data
  - Report all results, no matter their direction or statistical significance

Tools

protocols.io

bio-protocol

Penelope

- Protocols.io: open access repository of science methods; free to read & publish
- Bio-protocol: peer-reviewed protocol journal; free to read & publish
- Penelope: check your manuscript for reporting guideline compliance

Resources

equator network

SPIRIT

TIDieR
Template for Intervention Description and Replication

CONSORT

- Equator network (database of reporting guidelines): www.equator-network.org/
- Minimum set of items to address in protocol + report
- Study design specific
- Developed through consensus by stakeholders
- Evidence-based to contain bias, maximize transparency, & maximize utility

-https://github.com/mccartma/USGS_Consumptive_Use

Main Takeaways

Readability>Optimization

Keep it simple

Don’t write code if tired or in bad mood

Automation is main goal

Use relevant/quality names for variables

Use google!!!!!! Stack Overflow is your best friend

Know how your code works--comment comment comment

You have to start somewhere! So don’t worry if you aren’t the world’s best coder, you’ll get there :)

Questions?

goo.gl/ncBnr2