



August 21, 22, and 23, 2019, Harvard University, Cambridge

Welcome

Mercè Crosas, Ph.D.

Chief Data Science and Technology Officer, IQSS
Harvard University's Research Data Officer, HUIT
@mercecrosas

About a year ago,
reproducibility caught the
eye of *even* Congress

“Concerns about reproducibility and replicability have been expressed in both scientific and popular media. As these concerns came to light, Congress requested that the National Academies of Sciences, Engineering, and Medicine conduct a study to assess the extent of issues related to reproducibility and replicability and to offer recommendations for improving rigor and transparency in scientific research.”

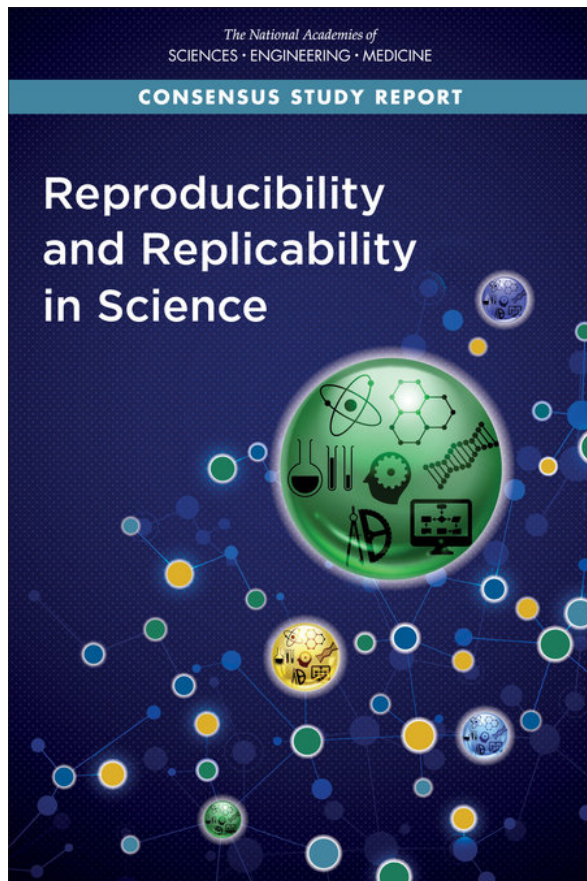
Reproducibility

Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.
[computational reproducibility]

Replicability

Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Definitions from the National Academies of Sciences, Engineering, and Medicine Consensus Study Report Highlights,
Reproducibility and Replicability in Science



Report Highlights include:

- **No crisis**, but we must do better
- Include a clear, specific, and complete description of how results are reached
- Promote use of open source tools
- Ensure computational reproducibility during peer review
- Facilitate transparent sharing and availability of digital artifacts, such as data and code

Published May 2019: <http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm>

More than a decade ago,
reproducibility had already
caught the eye of IQSS.

We developed the Dataverse project, an open-source platform to:

“facilitate transparent sharing and availability of digital artifacts, such as data and code”

Data (+Code) Sharing with Dataverse

- Data citation with a persistent identifier (DOI)
- Standard metadata, plus custom metadata
- Tiered access to data:
 - Fully Open, CC0
 - Register to access; Guestbook
 - Restricted with DUA
- Multiple versions of a dataset
- FAIR principles (Findable, Accessible, Interoperable, Reusable)

The image shows a screenshot of the Harvard Dataverse website. The header includes the Harvard Dataverse logo and navigation links like 'Search', 'Help', 'Guides', 'Support', and 'Sign Up'. The main content area features a search bar with the text 'Search over 90,200 datasets...' and a 'Find' button. Below the search bar, there is a 'Browse by subject' section listing various academic fields with their respective dataset counts. Overlaid on this interface are large, bold text statistics in red and orange, and a central black text block.

Harvard Dataverse (dataverse.harvard.edu):

- 90,000 datasets
- 500,000 files
- 8 million downloads
- 2,800 datasets in biomedical and life sciences

+ 45 other Dataverse sites (dataverse.org)

hosted across 6 continents

Subject Counts:

Subject	Count
Agricultural Sciences	1,088
Arts and Humanities	805
Astronomy and Astrophysics	521
Business and Management	422
Chemistry	183
Computer and Information Science	937
Earth and Environmental Sciences	1,359
Engineering	119
Law	277
Mathematical Sciences	213
Medicine, Health and Life Sciences	2,876
Physics	516
Social Sciences	38,502

Reproducible?

8,000 of the 90,000
datasets in Harvard
Dataverse contain the
files to reproduce the
published results.

Documentation

Data

R Code

HARVARD
Dataverse

Search ▾ About User Guide Support Sign Up Log In

Virus Epidemiology and Control (VEC) Dataverse (Kemri Wellcome Trust Research Programme, Kilifi, Kenya) Population dynamics of viral pathogens informing intervention strategies

Replication Data for: Whole genome sequencing and phylogenetic analysis of Human metapneumovirus strains from Kenya and Zambia Version 1.0

Kamau, Evelyn; Nokes, David James, 2019, "Replication Data for: Whole genome sequencing and phylogenetic analysis of Human metapneumovirus strains from Kenya and Zambia", <https://doi.org/10.7910/DVN/TVB65V>, Harvard Dataverse, V1 [Cite Dataset ▾](#) [Learn about Data Citation Standards.](#)

Description Replication dataset for plotting Figure 2 and Supplementary Figure 2: The average pairwise identities were calculated using the sequence data described in the article in Geneious R8.1.5 (<https://www.geneious.com>). The plots were done in R 3.5.3.

Subject Medicine, Health and Life Sciences

Keyword Human metapneumovirus, Whole genome sequencing, phylogenetic analysis

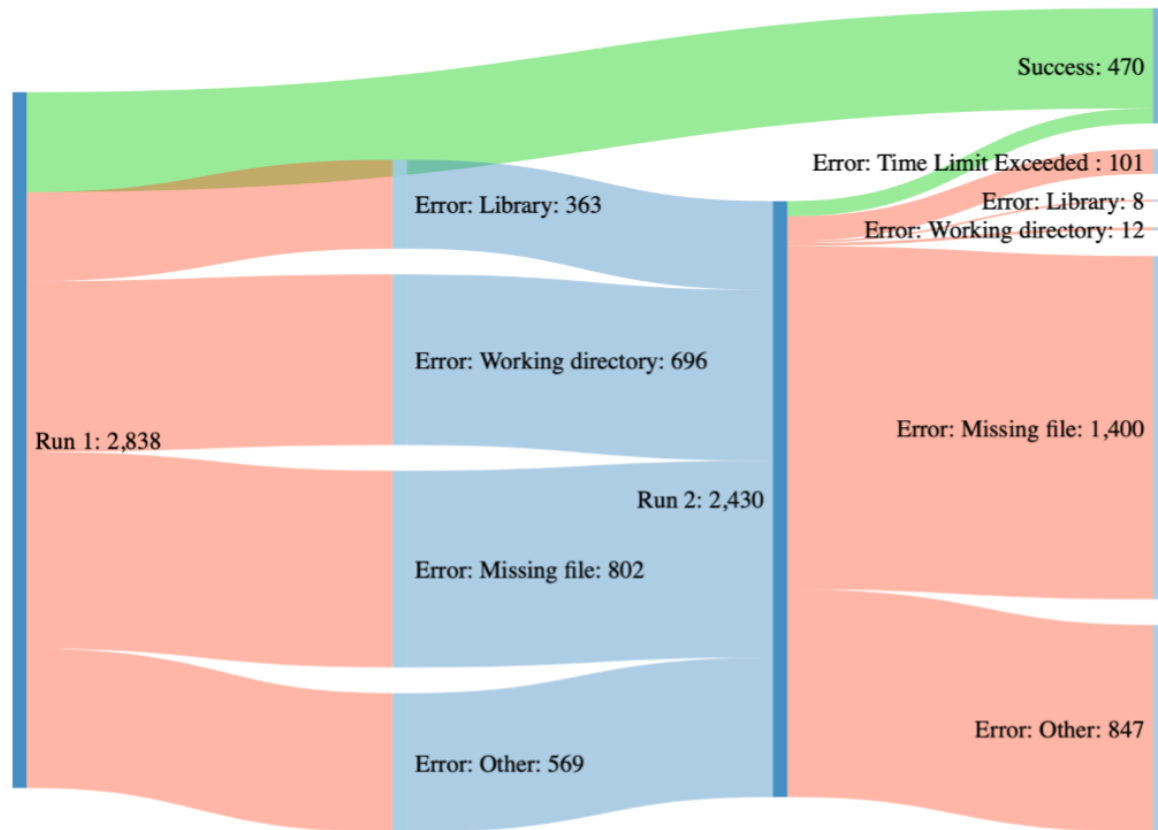
Files Metadata Terms Versions

1 to 6 of 6 Files [Download](#)

	EKamau_HMPV_WGS_Readme.txt Plain Text - 4.5 KB - Aug 5, 2019 - 0 Downloads MD5: 94e1f85ded8a0a8b4e99f460ba7de65f Dataset readme file Documentation	Download
	Identity_graph_HMPVA_Ggene.csv Comma Separated Values - 3.2 KB - Aug 5, 2019 - 0 Downloads MD5: 85b9d82a093f56f425a618da56dbba64 Data	Download
	Identity_graph_HMPVA_SHgene.csv Comma Separated Values - 2.4 KB - Aug 5, 2019 - 0 Downloads MD5: 5eec8e812e0c9cdd1a81e7d31a7cf551 Data	Download
	Identity_graph_HMPVB_Ggene.csv Comma Separated Values - 3.6 KB - Aug 5, 2019 - 0 Downloads MD5: 991131141a43d62276cd3083fd78a7d9 Data	Download
	Identity_graph_HMPVB_SHgene.csv Comma Separated Values - 2.6 KB - Aug 5, 2019 - 0 Downloads MD5: c8a3d807c5e88443678bfc3b68291802 Data	Download
	script_2Jul2019.R R Syntax - 3.0 KB - Aug 5, 2019 - 0 Downloads MD5: 64531365d4f6caaeaf95549d170fdcd Replication code in R Code	Download

But, it's complicated.

85.6% of archived R-based studies are not easily re-executable



From: Coding Better: Assessing and Improving the Reproducibility of R-Based Research with containR by Chris Chen

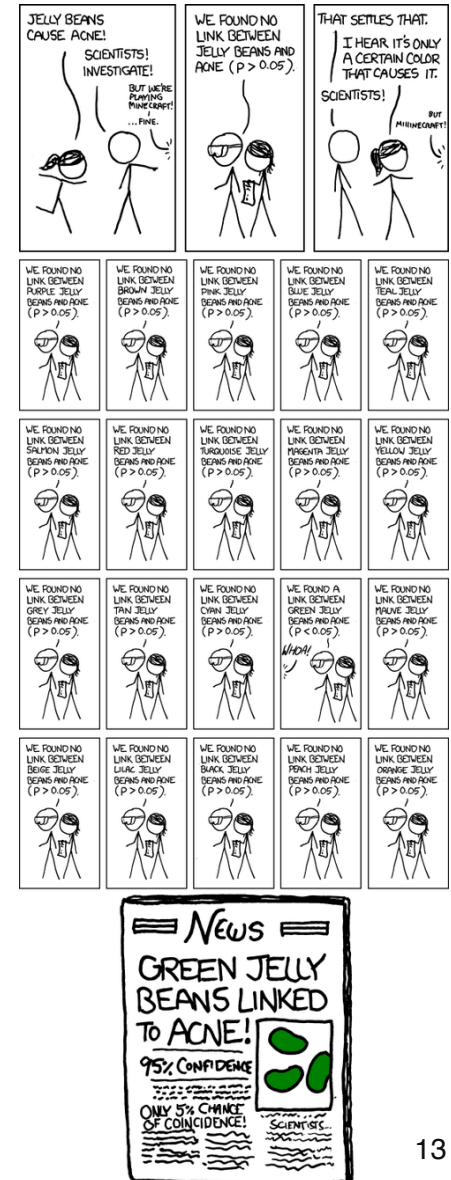
Can we do better?

Current Dataverse projects to improve Computational Reproducibility

- Include [reproducibility as part of the peer-review](#) workflow in journal dataverses
- Integrate with reproducibility and computational web-based tools (e.g., Code Ocean) to [facilitate code execution](#)
- Publish a [capsule](#) (container with data and code) verified for reproducibility
- When possible, [automate code execution](#) upon depositing the data and code

Computational Reproducibility is not sufficient

- Be mindful of publication bias and specification searching (statistical power, p-values, effect size)
- Include a detailed description of the analysis:
 - all methods, instruments, materials, procedures;
 - decisions for the exclusion or inclusion of data;
 - the analytic decisions and when these decisions were;
 - a discussion of the expected constraints on generality
 - reporting of precision or statistical power; and
 - discussion of the uncertainty of the measurements, results, and inferences;
- Conduct meta-analysis for replicability



NASEM, 2019, Reproducibility and Replicability in Science

& Christensen, Freese, Miguel, 2019, Transparent and Reproducible Social Science Research

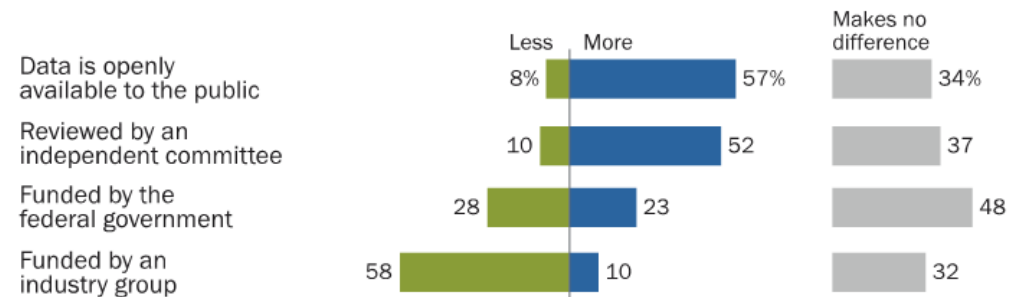
"Americans say open access to data (+ code) and independent review inspire more trust in research findings"

Trust and Mistrust of American Views on Scientific Experts, Pew Research Center, August 2, 2019

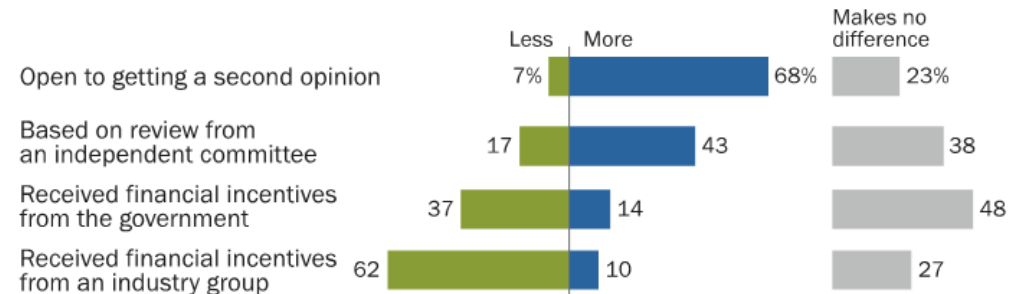
*(+ code) not in original statement

Majority of Americans say they are more apt to trust research when the data is openly available

% of U.S. adults who say when they hear each of the following, they trust scientific research findings ...



% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation ...



Note: Respondents who did not give an answer are not shown.

Source: Survey conducted Jan. 7-21, 2019.

"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER