Introduction to Open Scholarship

and Overview of Love Data Week

Chen Chiu

Website: dataservices.library.jhu.edu

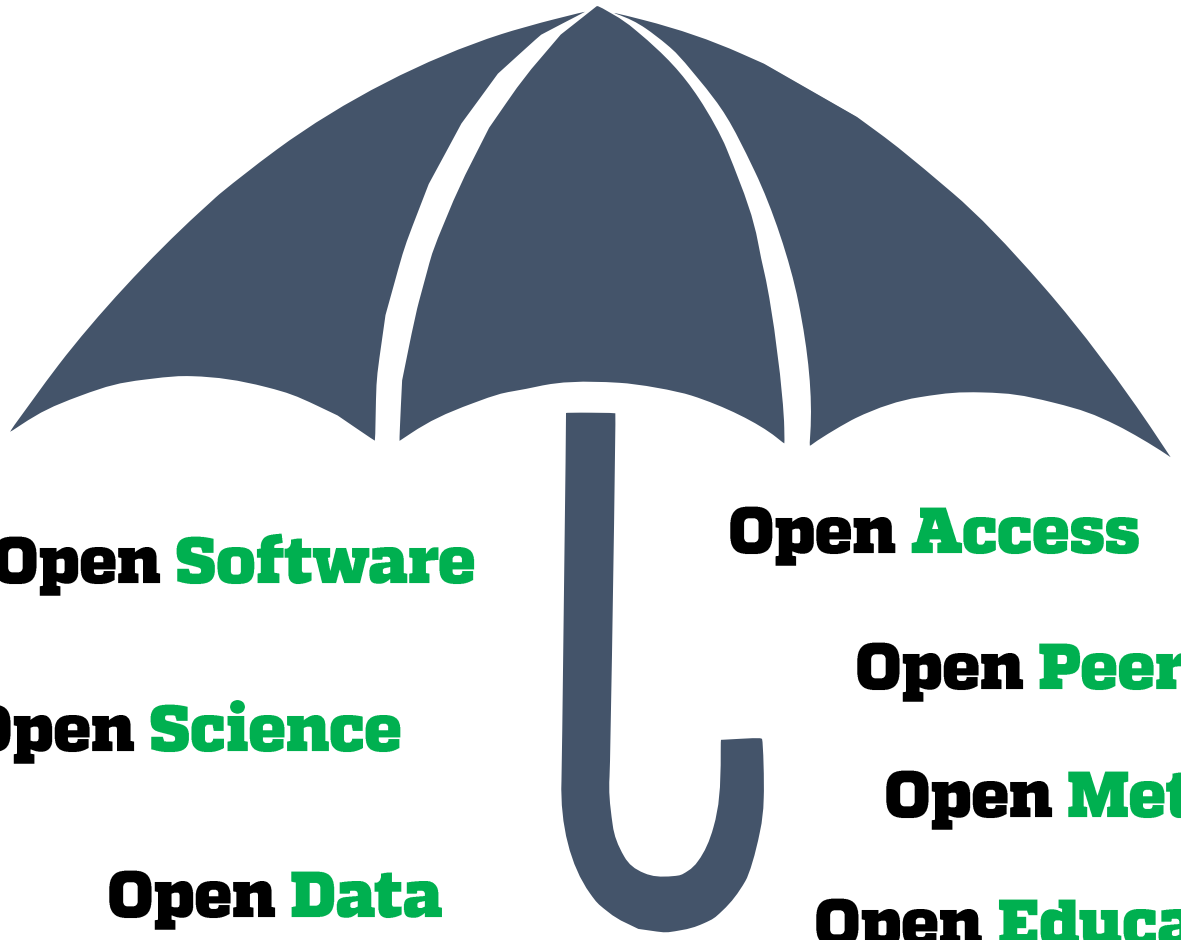
Email: dataservices@jhu.edu

Johns Hopkins Research Data Repository: archive.data.jhu.edu

JOHNS HOPKINS
LIBRARIES

Data Services

What is Open Scholarship?



Open Software

Open Science

Open Data

Open Access

Open Peer Review

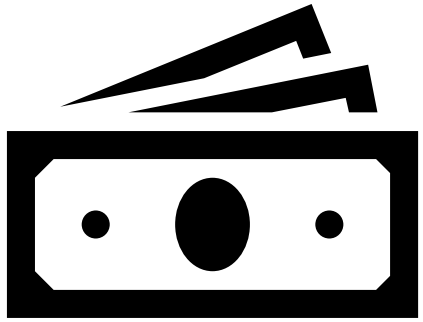
Open Methodology

Open Education Resources

- An umbrella term for all forms of openness of scholarly products
- A change of culture from traditional scholarly outputs

Traditional Scholarly Outputs

Funding

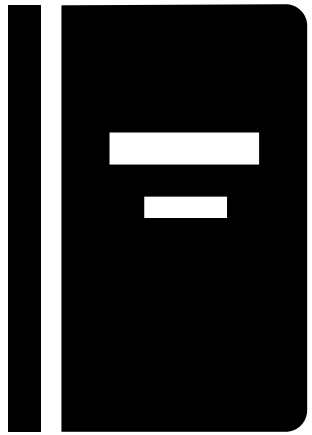


Scholarly Outputs

- Journal articles
- Conference presentations
- Original data
- Software/code
- Protocols
- Images/photos
- Videos
- Field notes
- Interview scripts

Most of the above are not shared

Publication



You cannot reproduce and validate the results by reading journal articles

Barriers to Open Scholarship

OPEN

means
reducing barriers
to accessing
scholarly products

Cost barrier

Solution: Make research products open and free access for the end users

Access barrier

Solution: Make research discoverable, understandable, and in a standard format

Reuse barrier

Solution: Researcher applies liberal reuse licenses for sharing and modifications

Funder Requirements: National Science Foundation

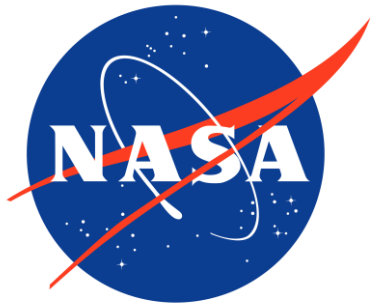
- First federal agency to request Data Management Plans
- NSF Open Access Plan 2.0
  - Persistent Identifiers (PIDs) for researchers: Required for senior personnel
  - Peer-reviewed publications: Made available on NSF Public Access Repository (PAR)
  - Data associated with a peer-reviewed publication: Share in discipline-specific repositories
  - Data NOT associated with a peer-reviewed publication: Guideline is under development
  - Software: No specific guidelines about sharing

Funder Requirements: National Institutes of Health

- Researchers are required to have ORCID (NOT-OD-19-109)
- Publication should be shared on PubMed Central for public access (Public Access Policy)
- Researchers are expected to maximize the appropriate sharing of scientific data (Policy for Data Management and Sharing)
  - Consider to share data in NIH domain-specific repositories first
  - Indicate whether specialized software/tools/code are needed to process or manipulate data
  - Having human participant data is not a justification for not sharing
  - Need to protect the privacy of participants while sharing data
  - More information on the Scientific Data Sharing website

Funder Requirements: National Aeronautics and Space Administration

- Persistent Identifiers are required for Investigators (ORCID is recommended)
- Sharing research outputs:
  - Materials from science events, such as workshops, conferences, etc.
  - Make publication publicly accessible on the NASA's repository
  - Research data
    - Should share in NASA archive
    - Use a generalist repository only if no domain-specific one available
    - Should have a CC0 license
  - Both data and software
    - Can be shared as supplementary materials of journal articles or in a public repository
    - Should be citable (GitHub is not sufficient)



Resource: NASA's Open Science Guideline

Summary of Funder Requirements

- Publishing papers in journals is not enough
- Having persistent identifiers for researchers
- Making publications open access
- Sharing research data in public repositories
- Sharing software is not required but recommended

Publisher Requirements: PLOS

- *PLOS* journals require authors to
  - Make all data **necessary to replicate** their study's findings publicly available **without restriction** at the time of publication
  - When specific **legal or ethical restrictions** prohibit public sharing of a data set, authors must indicate how others may obtain access to the data

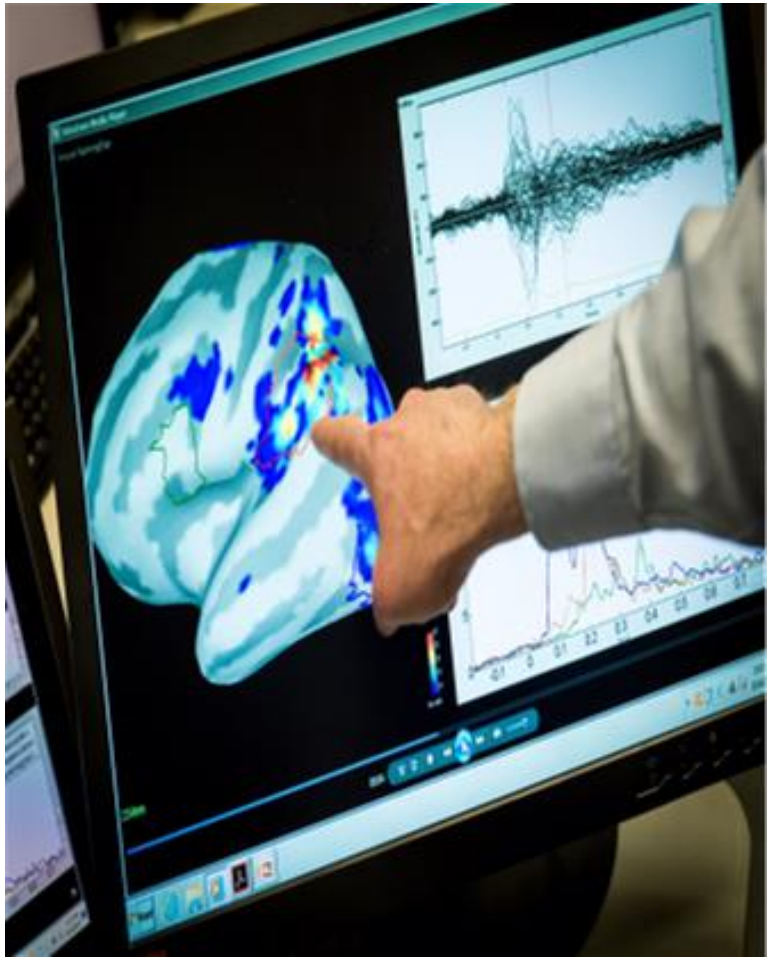
Publisher Requirements: Science

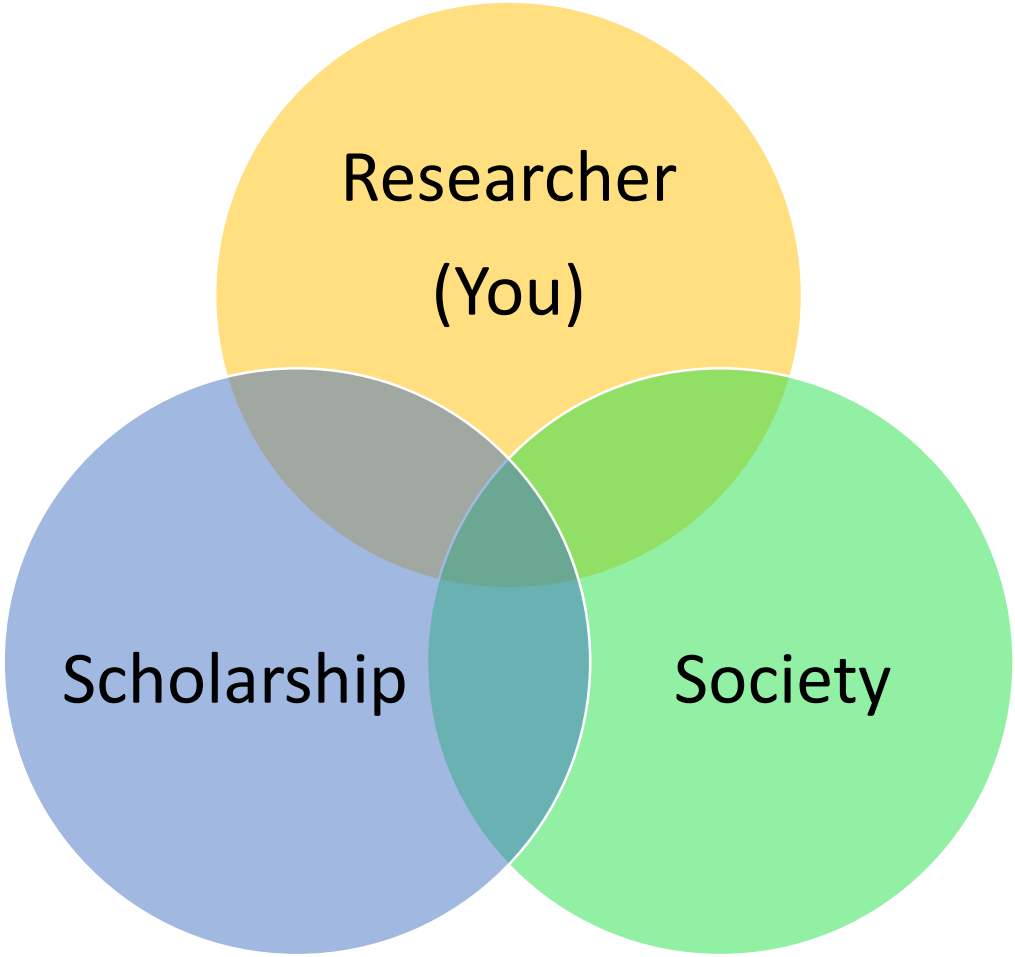
- The *Science* journals generally require
  - **All data underlying the results** in published papers to be publicly and immediately available
  - Post-publication **embargoes are not permitted**, nor are stipulations for readers to contact the authors

Science

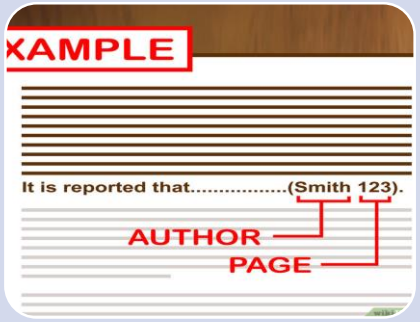
AAAS

Benefits of Open Scholarship





Benefits to You



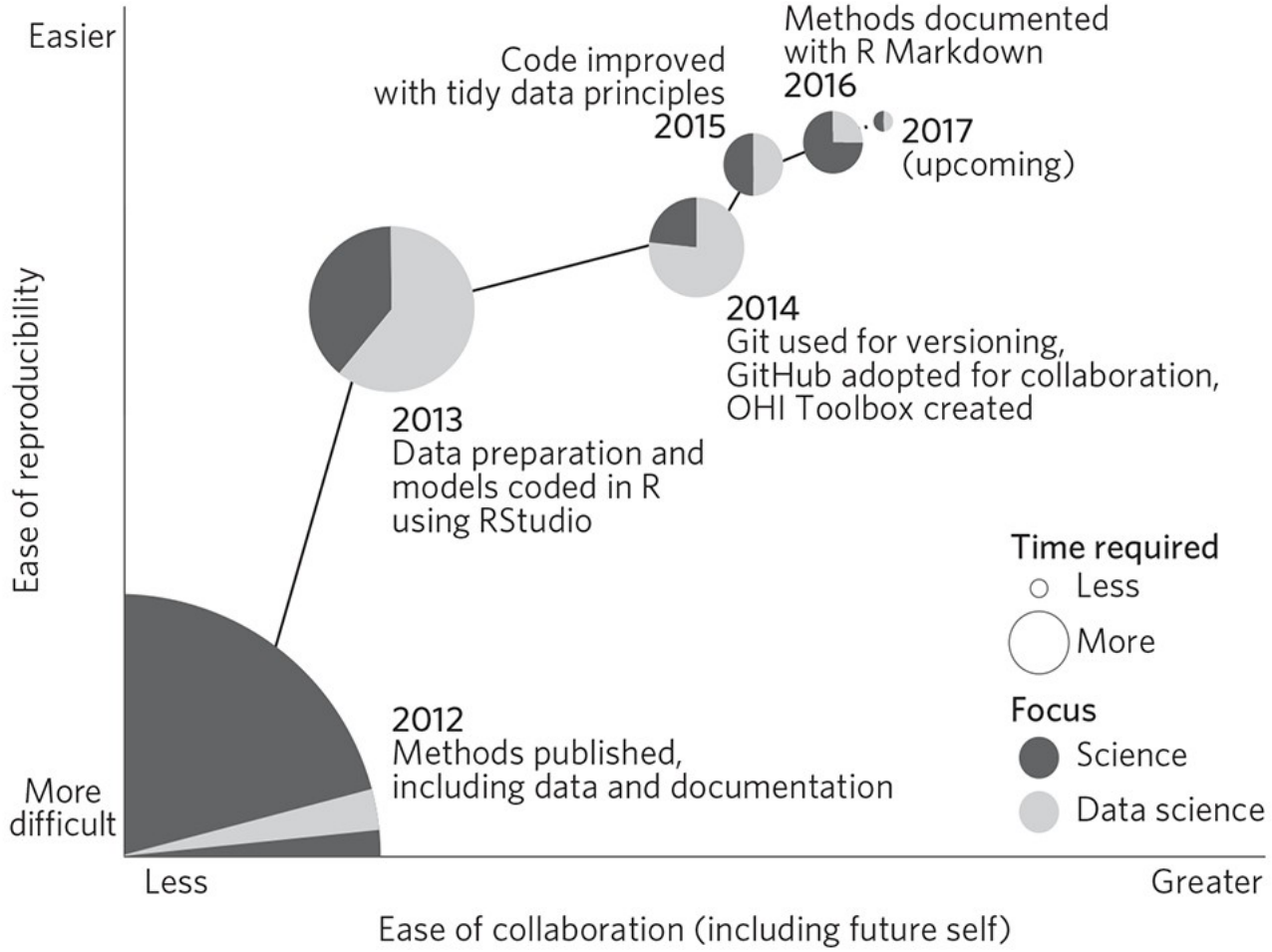
Increase article citation

Comply to funders' and publishers' data sharing requests

Improve data quality

Conduct research more efficiently

Example of Improved Efficiency



"Increased reproducibility and collaboration has reduced the amount of time required to repeat methods (size of bubbles) with updated data annually, allowing us to focus on improving methods each year (text labels show the biggest innovations)."

Lowndes, J., Best, B., Scarborough, C. et al. Our path to better science in less time using open data science tools. Nat Ecol Evol 1, 0160 (2017). https://doi.org/10.1038/s41559-017-0160

Benefits to Scholarship

Solve

complex problems

Improve

peer evaluation of research and scientific rigor

Reduce

publication bias

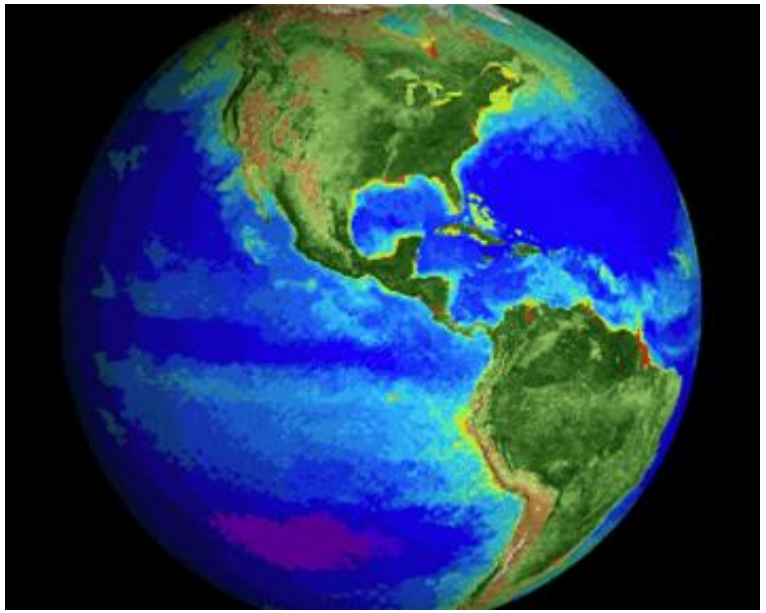
Solve Complex Issues

Open Scholarship helps address complex scientific questions, such as climate change

Temperature

Climate model code

Photosynthetic output



Atmospheric CO_2

Oceanic pH

Area covered by polar ice caps

This satellite data lets scientists watch Earth "breathe" and improves our understanding of our home planet. Someday, this same satellite technology could help in the search for life on other planets.

Improve Evaluation and Rigor of Research

- **Encourages the transparency** of the entire research lifecycle (e.g., sharing of methods, raw data, code, documentation, peer review)
- **Facilitates reproducibility** and replication
- Promotes study pre-registration to **reduce HARKing and p-hacking**
- **Increases statistical power** of studies and standardization by encouraging team science and meta-analyses

Munafò, Marcus R. A manifesto for reproducible science. Nature Human Behaviour volume1, Article number: 0021 (2017) doi: 10.1038/s41562-016-0021.

Reduce Publication Bias

Publication bias toward positive and novel results

- **Pre-register** your studies
- Post a **pre-print**
- Publish in a journal receptive to non-novel work
- Submit a **registered report**

Benefits to Society

Industries

Policymakers

NGOs

Citizen Scientists

Concerns for Practicing Open Scholarship

Of course I want to further science by sharing all my research products, BUT I am worried about …

No money!

Scooping?

No time!

Misuse?

Re-identification?

Scrutiny?

Concern: Scooping

You can:

- Preregister your experiment so that your ideas are public and with a timestamp
- Post your pre-prints prior to formal peer review
- Embargo the public sharing of your research for a limited time
  - Caution: Some funders/publishers may NOT allow an embargo
- Submit to a journal that allows complementary studies

Concern: Misuse

You can:

- Protect yourself by making data, documentation, protocols, code, etc. open
- Document the intended purpose of your dataset
- Place some restrictions on sharing and reuse

Concern: No Time

You can:

- Organize and document your scholarly outputs from the beginning of your project to save time at the end
- Review the requirements of your chosen repository for depositing your research to ensure deposit goes smoothly

Concerns: No Money

You can:

- Request money from your funder to cover these costs
  - Write data management and sharing costs into grant proposal budgets
- Look for funding to support open science
- Look for free and open-source tools

Concerns: Re-identification

You need to:

- Get de-identification review and approval to share from the Data Trust Council for JH Medicine data
- Have explicit consent from participants to share de-identified data in a repository
- Make sure that data are fully de-identified; otherwise, share in a repository with some restrictions

Concern: Scrutiny

Things to keep in mind:

- Keep good records of everything (e.g., experimental and data analysis procedures) and share these along with your data and code
- Ultimately, as open scholarship becomes part of everyone's normal workflow, it won’t be as scary

Example: NextStrain

- Winner of the 2017 NIH Open Science Prize
- An open-source project to harness the potential of pathogen genome data
- Track multiple viruses, including COVID-19 and Zika
- Provide the following scholarly products:
  - a continually updated view of publicly available data
  - powerful analytic and visualization tools
- Source code is freely available under a nonrestrictive license
- Benefits the community
  - Real-time tracking of Tomato brown rugose fruit virus (ToBRFV) outbreaks in the Netherlands using Nextstrain (https://doi.org/10.1371/journal.pone.0234671)
  - Molecular Epidemiology of HIV-1 in Eastern Europe and Russia (https://doi.org/10.3390/v14102099)

Nextstrain

https://nextstrain.org/

Example: OpenTopography

- Winner of the AGU Advancing Earth and Space Sciences Open Science Recognition Prize 2023
- Facilitates efficient access to topography data, tools, and resources to advance our understanding of the Earth's surface, vegetation, and built environment
- Data are free and open access
- Offers workshops and online tutorials
- Partner with them to receive
  - hosting, managing, distributing, and processing of high resolution topography
  - on-demand tools for processing data
  - big data topography analysis

OpenTopography
High-Resolution Topography Data and Tools

https://opentopography.org/

Example: eBird

- A citizen science project
- The world's largest biodiversity-related science projects
- Gather information from birders and freely share it
- Power new data-driven approaches to science, conservation, and education
- Birders enter observations via a mobile app with quality control for data entries
- Tools available freely to analyze data
- Benefits the community
  - BirdFlow: Learning seasonal bird movements from eBird data (https://doi.org/10.1111/2041-210X.14052)
  - Using eBird data to model population change of migratory bird species (https://doi.org/10.5751/ACE-00960-120104)

https://ebird.org/home

What Makes an Open Scholarship Project Successful?

- A public web platform to host, manage, submit, and access data
- A platform that is maintained, improved, and updated
- Data are open and free to access
- Software tools are available to process, analyze, or visualize data
- Good documentation or tutorials available to show how to use these data and tools
- Data is entered following quality control measures to ensure data are standardized

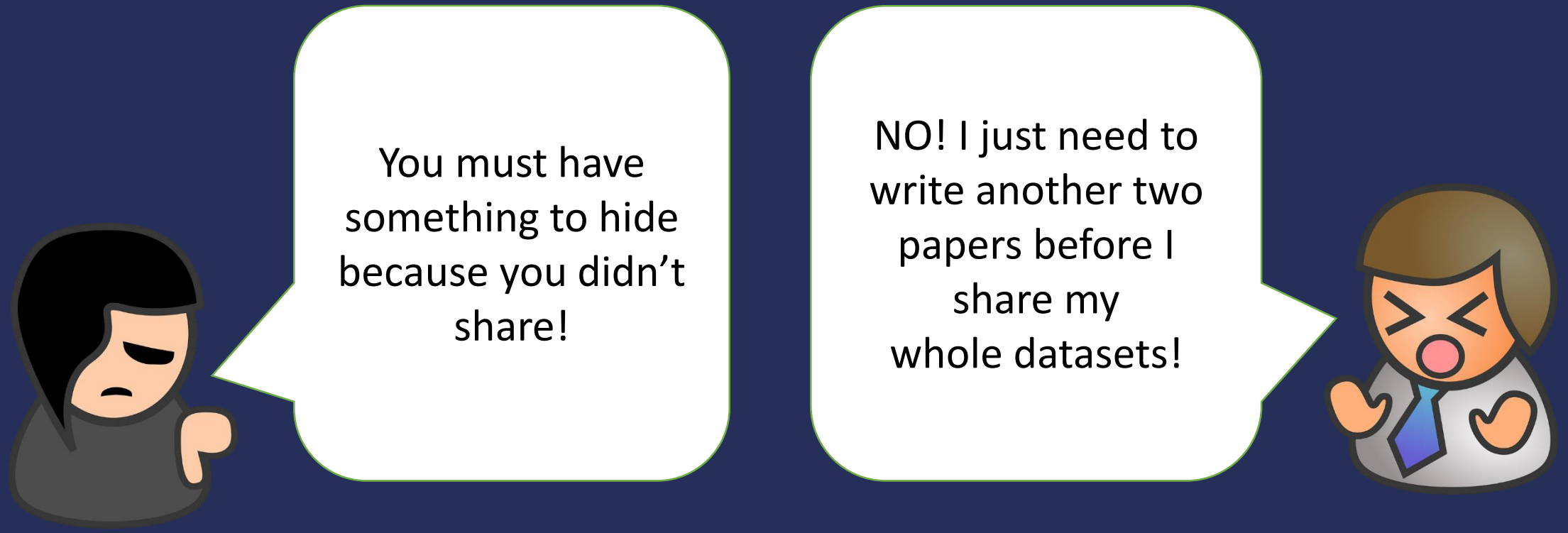
Making scholarly products open is not enough, you need to make research reproducible

Provide Enough Documentation to Make Research Reproducible

- **Equipment:** Information about equipment and experimental setup
- **Procedures:** Steps to conduct your experiment(s)
- **Data/code sources:** List sources for data and code (if you are not the original owner)
- **Data analysis:** Methods and steps to conduct data cleaning and analysis
- **Documentation:** Information that is necessary to understand your scholarly products
- **Persistent location:** A location where others can always link to and find your research products
- **License:** Terms for people to use your scholarly products

MYTH

Those scientists who do not share data/code must have something to hide!



A Workshop to Reproduce Results in Published Papers

- NLM Reproducibility Workshop: Held in 2019 by National Library of Medicine
- Lessons learned from this workshop:
  - Task: 5 teams tried to reproduce bioinformatics papers
  - Results: No papers were successfully reproduced

Conclusion

It’s not easy to reproduce a research project

- Missing data or code or analysis tools
- Hard to follow workflow
- Versioning issues
- Hard to reproduce the computing environment

BUT…

Some teams reached out to corresponding authors and got quick responses!

Most authors were willing to work with them to make it possible to reproduce their results.

Reproduce Other People's Results is NOT an easy task!

http://m.atchuup.com/pinterest-baking-fails/

https://www.boredpanda.com/funny-pinterest-fails/?utm_source=google&utm_medium=organic&utm_campaign=organic

Reproducible Research Examples

We built multiple hierarchical models to examine how increased temperature and other factors influenced shrub seedling recruitment, growth and mortality, as well as tussock grass gap dynamics. For each model, we used Bayesian inference and fitted models in r 3.3.2 (R Core Team, 2016) using package rstan 2.14.1 (Stan Development Team, 2016). Detailed information about experimental design and analysis is provided in Supporting Information. Data and source code are available at: https://github.com/jscamac/Alpine_Shrub_Experiment . To aid in the reproducibility of this work, our code was written using a remake framework (FitzJohn, 2015). This allows others to readily reproduce our entire workflow from data processing, through to producing a pdf of this manuscript by calling remake::make. To safeguard against cross-platform issues and future software changes, we have embedded this framework within a Docker image (https://hub.docker.com/r/jscamac/alpine_shrub_experiment).

~ Global Change Biology (2017) 23, 3249-3258, doi: 10/1111/gcb.13614

This is a good example of reproducible research. Why?

Reproducible Research Examples

This is a good example of reproducible research. Why?

- It has data and source code available on GitHub
- It has a paragraph explaining the effort they spent on making this study reproducible
- This article is open access
- All tools they used are open source
- There is a license attached to the source code
- A well-written README

What Can be Improved?

- GitHub doesn’t provide a persistent identifier for your code
- Why is this a problem?
  - You may one day decide to reorganize your GitHub repositories so this URL will change
  - People who use your journal article link to your code will get a broken link
- **Solution: Assign a persistent identifier for your source code or data**
- **Benefits** of persistent identifiers: A persistent link so people can always find your code or data

Code Reproducibility

- Many scientists write code to acquire or analyze data, but
  - Don’t have formal training in computer science
  - Are self-taught programmers
  - Only have time to make the code work
  - Not enough time or knowledge to make the code efficient, clean up scripts, write good documentation

Best Practices for Documenting Code

Tip #1: Give meaningful names to variables and functions

Tip #2: Write helpful comments and use a computational notebook to document code

Tip #3: Include a ReadMe file

Tip #4: Use a version control system to track changes

Tip #5: Document the environment and dependencies

Tip #6: Share code on an online code repository

To learn more, check out our
Documenting Your Research Data online modules

Pick 2-3 items to start your open scholarship journey

Pick one thing from the list that you can start doing soon:

- ☑ Obtain an ORCID ID
- ☑ Pre-register your study
- ☑ Use open source application for your study
- ☑ Provide good documentation for your scholarly products
- ☑ Use nonrestrictive licenses for your scholarly products
- ☑ Publish a preprint
- ☑ Share your scholarly products in a trusted repository
- ☑ Publish on a journal provides open access

Love Data Week Overview

Love Data Week is an international celebration of data.
This year's theme is "My Kind of Data".

Join us for a week-long series featuring online talks, panels, and workshops exploring how people participate in or are affected by data.

Topics include:

- How to find and use data in your research
- Ethical data sharing, management, and visualization
- Open Scholarship and Reproducible Research
- And much more!

Register here

https://bit.ly/jhu-love-data-week

LOVE DATA WEEK

SCHEDULE OF EVENTS

Learn more and register at: bit.ly/jhu-love-data-week

JOHNS HOPKINS LIBRARIES

MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
10 - 11 AM Keynote ★ **Environment, Health and Justice** *Dr. Ana Navas-Acien*	**10 - 11 AM Webinar** **Introduction to Open Scholarship and Overview of Love Data Week** *JHU Data Services*	**10 - 11 AM Workshop** **Navigating HathiTrust wtih Python** *Hale Sirin,* *Center for Digital Humanities*	**10 - 11 AM Session Talks** **An Introduction to Finding Data across Multiple Disciplines** *Sheridan Library and Welch Medical Library*	**10 - 11 AM Keynote** ★ **Science in the Era of AI** *Dr. Alexander S. Szalay*
11 AM - 12 PM Session Talks **Ethic in Data Visualization** *Alice Feng and Nancy Shin*	**11 AM - 12 PM Workshop** **Reproducible Research with Quarto and Jupyter Lab** *JHU Data Services*	**11 AM - 12 PM Workshop** **Exploring Archives with OCR (and LLMs)** *Sam Backer* *Center for Digital Humanities*	**11 AM - 12 PM Workshop** **Version Control: Using Git and GitHub** *JHU Data Services*	**11 AM - 12 PM Session Talks** **Analyzing the Archive: Working with Data from Text, Audio, and Video** *Kevin Hawkins, Casey Lurtz, Alexandre White*
	1 - 2 PM Workshop **Data for Studying the Quality of Elections and Democracy** *Bryce Corrigan, SNF Agora Institute*	**1 - 2 PM Workshop** **Organizing and Sharing Qualitative Data via the Qualitative Data Repository (QDR)**	**1 - 2 PM Session Talks** **Open Scholarship Support at JHU** *Data Services, OSPO, and IDIES/SciServer*	
2 - 3:30 PM Panel Discussion **The challenge of balancing biomedical data sharing and patient privacy protection at Johns Hopkins**	**2 - 3:30 PM Panel Discussion** **Open Scholarship at Johns Hopkins: Experiences and Advice from our Researchers**	**2 - 3:30 PM Panel Discussion** **Critical Approaches to Data Sharing: Justice Focused Initiatives**	**2 - 3:30 Workshop** **Designing Effective Data Visualizations** *JHU Data Services*	**2 - 4pm Workshop** **SciServer.org: Easy Online Analysis of Big Data for Research and Education** *IDIES*

Workshop:

Reproducible Research with Quarto and JupyterLab

Tuesday, February 13th, 11 am -12 pm

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Lubov McKone
Data Management Consultant
Johns Hopkins University Data Services

Pete Lawson
Data and Visualization Librarian
Johns Hopkins University Data Services

quarto

jupyter

Register Here

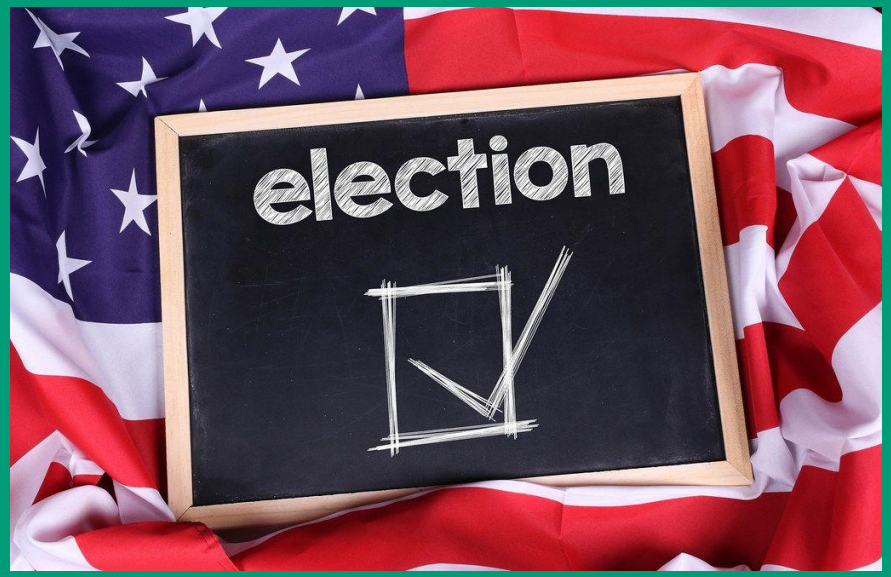
Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

Data for Studying the Quality of Elections and Democracy

Tuesday, February 13th, 1 – 2 pm

Bryce Corrigan
Statistician and Lecturer, SNF Agora Institute, Johns Hopkins University



LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Panel Discussion:

Open Scholarship at Johns Hopkins: Experiences and Advice from Our Researchers

Tuesday, February 13th 2:00 to 3:30 PM

LOVE DATA WEEK

Feb 12–16 | 2024

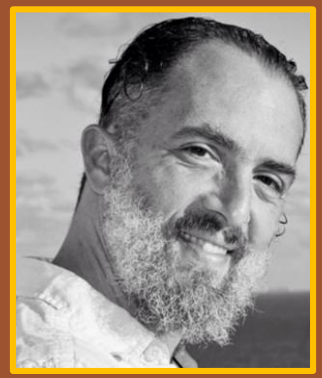
Register Here

Pete Aceves
Assistant Professor
Management and Organization
Carey Business School

Netz Arroyo
Associate Professor
Pharmacology
School of Medicine

Hunter Gehlbach
Professor
Education
School of Education

John Muschelli
Associate Scientist
Biostatistics
School of Public Health



Joshua Vogelstein
Associate Professor
Biomedical Engineering
School of Engineering

Lubov McKone
(Moderator)
Data Management Consultant
Data Services

Workshop:

Navigating HathiTrust with Python

Wednesday, February 14th, 10 – 11 am

Hale Sirin
Post-Doctoral Researcher
Johns Hopkins Center for Digital Humanities

HATHI TRUST DIGITAL LIBRARY

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

Exploring Archives with OCR (and LLMs)

Wednesday, February 14th, 11 am – 12 pm

Sam Backer
Postdoctoral Fellow
Johns Hopkins Center for Digital Humanities

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

Organizing and Sharing Qualitative Data via the Qualitative Data Repository (QDR)

Wednesday, February 14th, 1 – 2 pm

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Dessi Kirilova
Senior Curation Specialist
Qualitative Data Repository

QDR
The Qualitative Data Repository

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Panel Discussion:

Critical Approaches to Data Sharing

Wednesday, February 14th, 2 - 3:30 pm

LOVE DATA WEEK

Feb 12–16 | 2024

Register Here

Parker E. Foster
Research Associate
New York University Metro Center, Center for Policy, Research, and Evaluation

Sofia Locklear
Assistant Professor
University of Toronto Mississauga

Lauren Rubin
Director of Development
St. Francis Neighborhood Center

Kevin S. Hawkins
(Moderator)
Program Director for the Opioid Industry Documents Archive
Johns Hopkins University

Session Talks:

An Introduction to Finding Data Across Multiple Disciplines

Thursday, February 15th, 10 – 11 am

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Business Location Data via the Sheridan Libraries

Jim Gillispie

Social Science Librarian, Johns Hopkins University Sheridan Libraries

Data for Your Business Plan or Pitch

Katy Troeschel

Technology & Entrepreneurship Librarian, Johns Hopkins University Sheridan Libraries

A Brief Introduction to dbGaP: An Archive of Genotype/Phenotype Data from NCBI

Rob Wright

Basic Science Informationist, Johns Hopkins University Welch Medical Library)

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

Version Control: Using Git and GitHub

Thursday, February 15th, 11 am – 12 pm

Chen Chiu
Senior Data Management Consultant
Johns Hopkins University Data Services

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Session Talks:

Open Scholarship Support at JHU

Thursday, February 15th, 1 – 2 pm

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

An Overview of Data Services and the Institutional Repository

Chen Chiu

Senior Data Management Consultant

Johns Hopkins University Data Services

Introducing the Johns Hopkins Open Source Programs Office

Megan Forbes

Program Manager, Open Source Programs Office

Johns Hopkins University Sheridan Libraries

Introduction to the Institute for Data Intensive Engineering and Science, SciServer, and the Scientific Software Engineering Center

Gerard Lemson

Director of Science of The Institute for Data Intensive Engineering and Science and

Associate Director of the Scientific Software Engineering Center

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

Designing Effective Data Visualizations

Thursday, February 15th, 2 – 3:30 pm

Pete Lawson
Data and Visualization Librarian
Johns Hopkins University Data Services

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

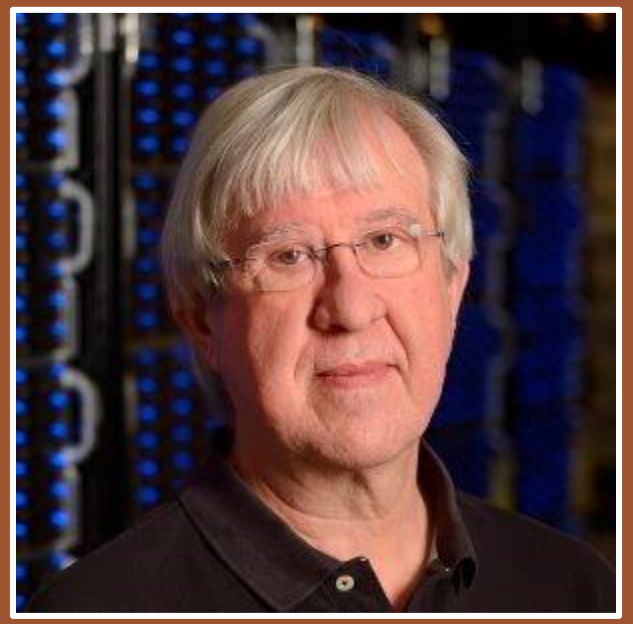
Closing Keynote Presentation

Friday, February 16th, 10 am -11 am

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!



Alexander S. Szalay
Professor of Physics and Astronomy
Johns Hopkins University

Science in the Era of AI

Artificial Intelligence is having a tremendous impact on every aspect of our life, including science. The talk will discuss various aspects of how science is changing very rapidly and contrasts trends in academia to those in the industry.

Register Here

Learn more and register at:
bit.ly/jhu-love-data-week

Session Talks:

Analyzing the Archive: Working with Data from Text, Audio, and Video

Friday, February 16th, 11 am – 12 pm

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Supporting use of computational methods on the Opioid Industry Documents Archive

Kevin S. Hawkins

Program Director for the Opioid Industry Documents Archive, Johns Hopkins University

Teaching with Historical Data

Casey Marina Lurtz

Assistant Professor of History, Johns Hopkins University

Underwriting Souls: Archives and Humanizing Data Digitizing Corporate

Alexandre White

Assistant Professor of Sociology, Johns Hopkins University

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Workshop:

SciServer.org: Easy Online Analysis of Big Data for Research and Education

Friday, February 16th, 2 – 4 pm

Jordan Raddick
Associate Director for Education at the Institute for Data-Intensive Engineering and Science (IDIES)
Johns Hopkins University

SciServer

Collaborative data-driven science

LOVE DATA WEEK

Feb 12–16 | 2024

Join us for a week-long celebration of data!

Register Here

Learn more and register at: bit.ly/jhu-love-data-week

Contact JHU Data Services

GO TO
dataservices.library.jhu.edu

EMAIL
dataservices@jhu.edu

SHARE DATA AT
archive.data.jhu.edu

Helping you

FIND

USE

MANAGE

VISUALIZE

SHARE

DATA

JOHNS HOPKINS LIBRARIES | Data Services