Carnegie
Mellon
University
Libraries

An End-to-end Open Science and Data Collaborations Program

Huajin Wang, Ph.D.

Melanie Gainey, Ph.D.

Co-directors, Open Science & Data Collaborations Program

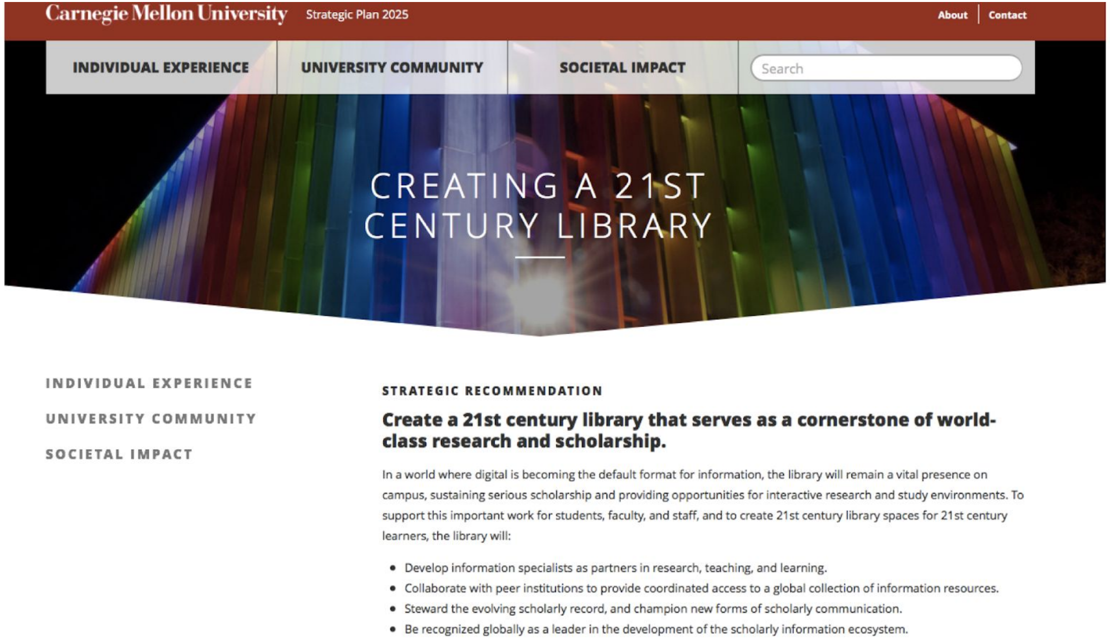
Carnegie Mellon University Libraries

December 7, 2021

Today's talk

1. Supporting open science at CMU and adapting to the pandemic
2. Developing instruments for impact assessment

How CMU Libraries Supports Open Science



CREATING A 21ST CENTURY LIBRARY

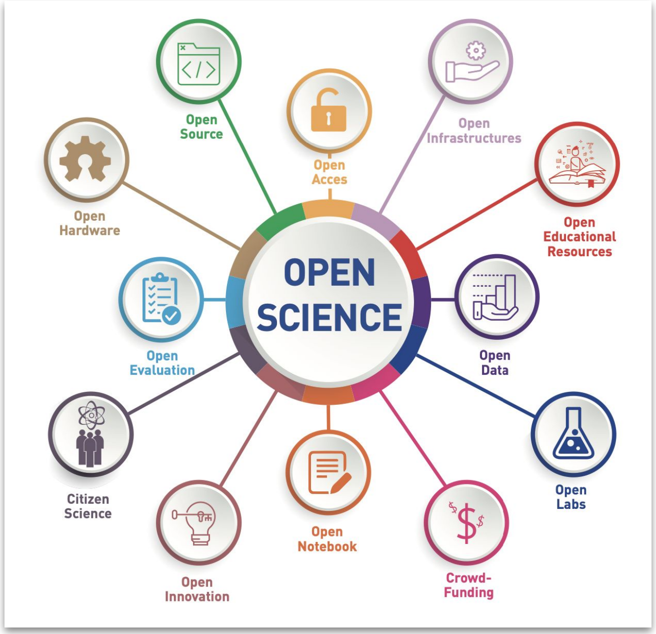
STRATEGIC RECOMMENDATION

Create a 21st century library that serves as a cornerstone of world-class research and scholarship.

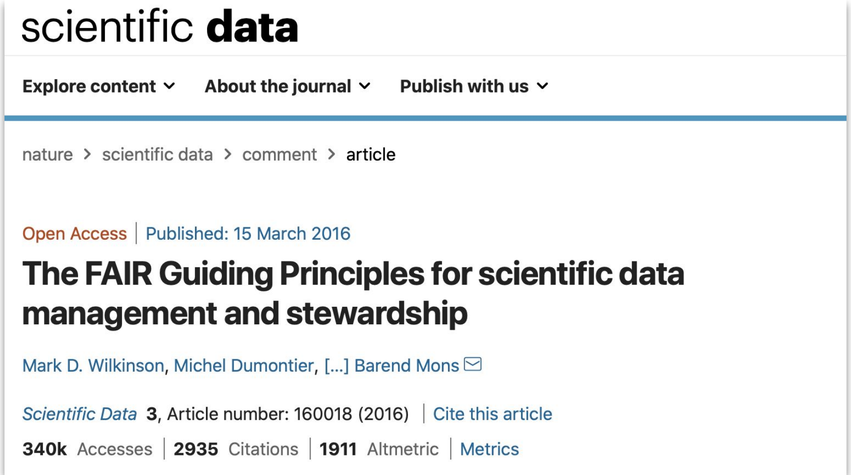
In a world where digital is becoming the default format for information, the library will remain a vital presence on campus, sustaining serious scholarship and providing opportunities for interactive research and study environments. To support this important work for students, faculty, and staff, and to create 21st century library spaces for 21st century learners, the library will:

- Develop information specialists as partners in research, teaching, and learning.
- Collaborate with peer institutions to provide coordinated access to a global collection of information resources.
- Steward the evolving scholarly record, and champion new forms of scholarly communication.
- Be recognized globally as a leader in the development of the scholarly information ecosystem.

What is Open Science / Open Research?



From: UNESCO UNESCO Recommendation on Open Science. 2021.


scientific **data**

nature › scientific data › comment › article

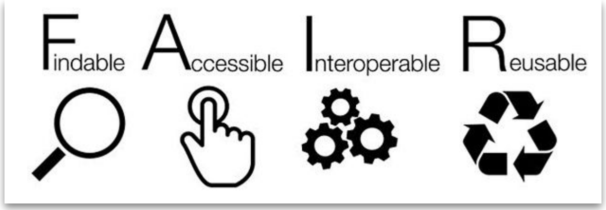
Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

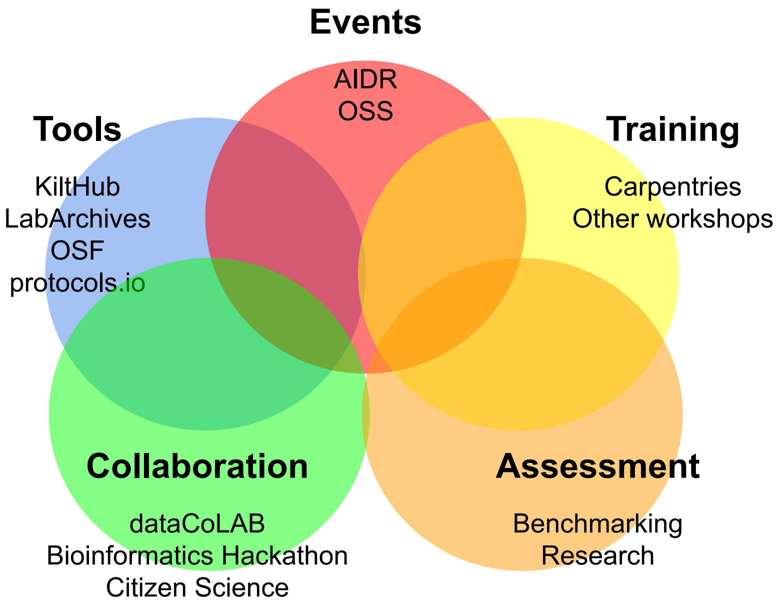
Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons ✉

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

340k Accesses | **2935** Citations | **1911** Altmetric | Metrics



Open Science & Data Collaborations Program



Events

AIDR
OSS

Tools

KiltHub
LabArchives
OSF
protocols.io

Training

Carpentries
Other workshops

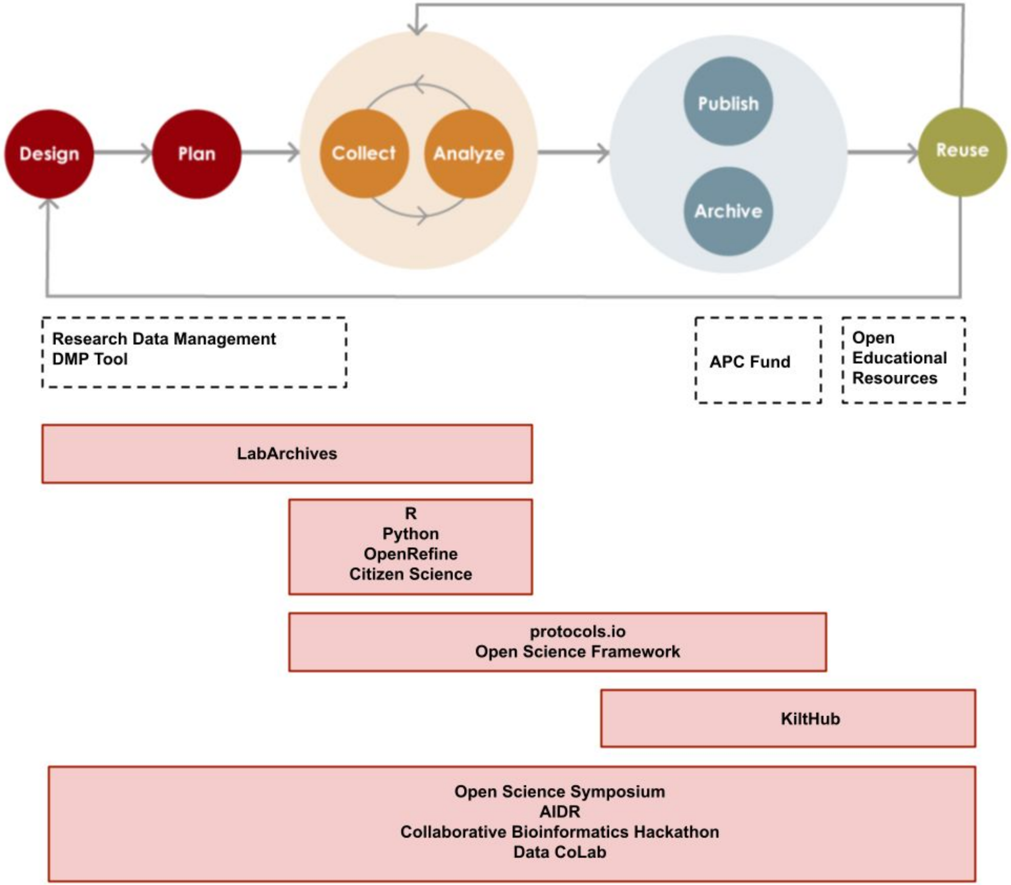
Collaboration

dataCoLAB
Bioinformatics Hackathon
Citizen Science

Assessment

Benchmarking
Research

https://www.library.cmu.edu/datapub/open-science



Research Data Management
DMP Tool

APC Fund

Open
Educational
Resources

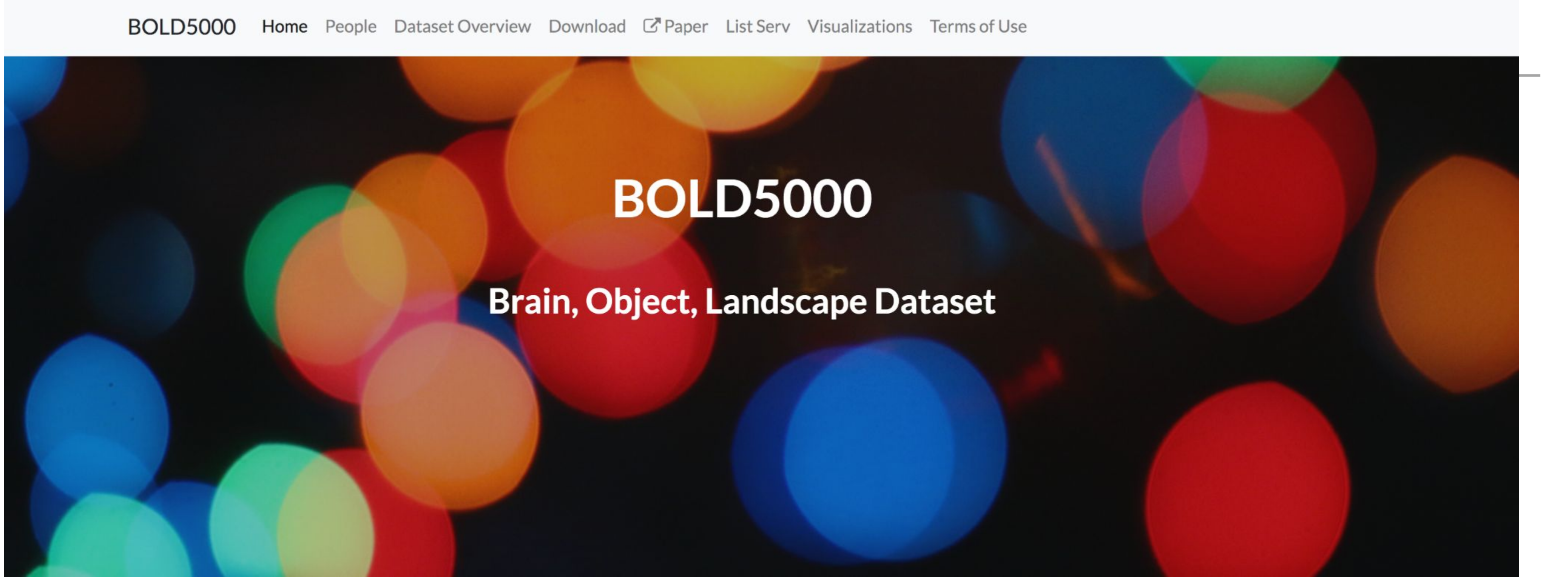
LabArchives

R
Python
OpenRefine
Citizen Science

protocols.io
Open Science Framework

KiltHub

Open Science Symposium
AIDR
Collaborative Bioinformatics Hackathon
Data CoLab

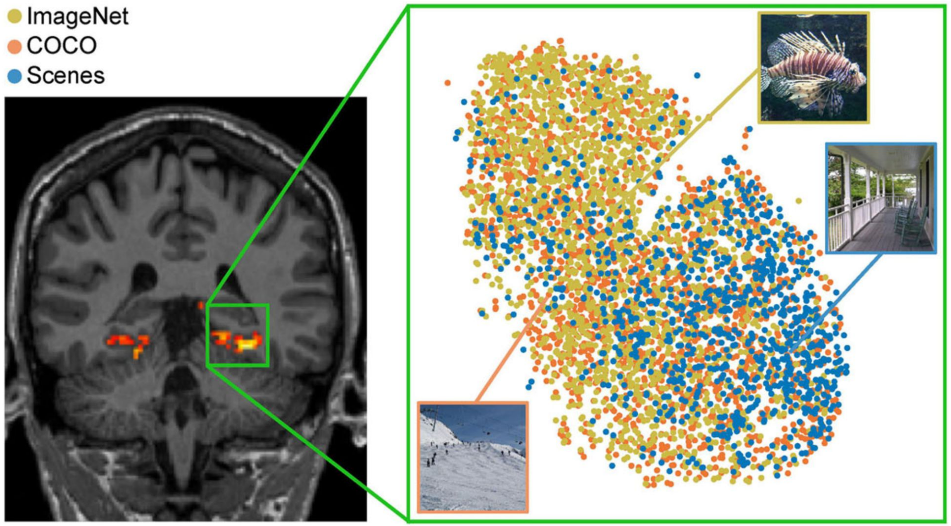


BOLD5000

Brain, Object, Landscape Dataset

bold5000.org

Largest slow event-related fMRI dataset



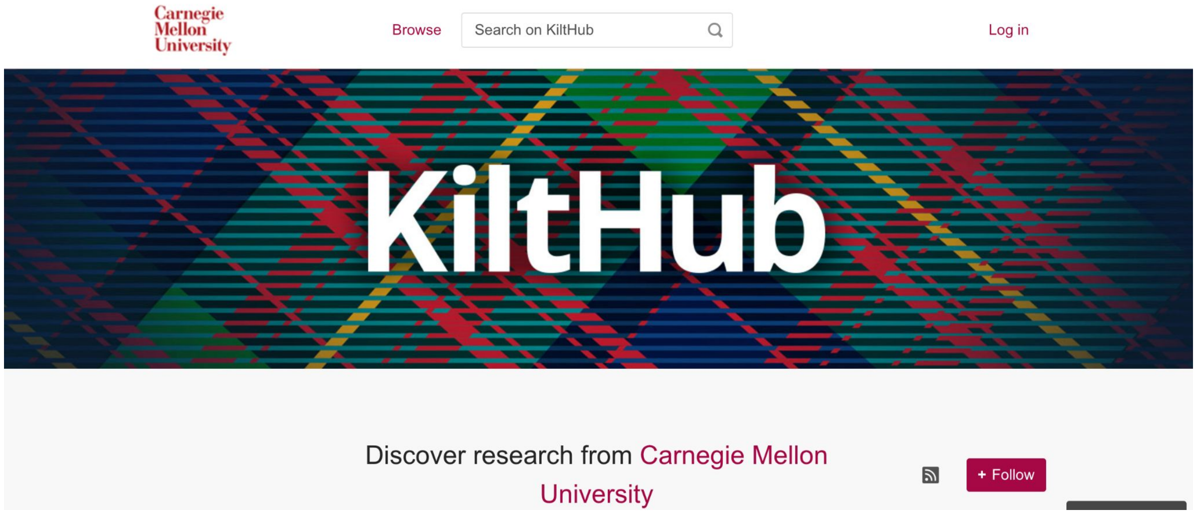
https://www.cmu.edu/news/stories/archives/2019/may/dataset-bridges-human-vision-machine-learning.html

Collaboration

Computer-vision scientists and visual neuroscientists essentially have the same end goal: to understand how to process and interpret visual information." - Nadine Chang, Robotics Institute, CMU

https://www.cmu.edu/news/stories/archives/2019/may/dataset-bridges-human-vision-machine-learning.html

Sharing data in a trusted repository



Carnegie Mellon University

Browse

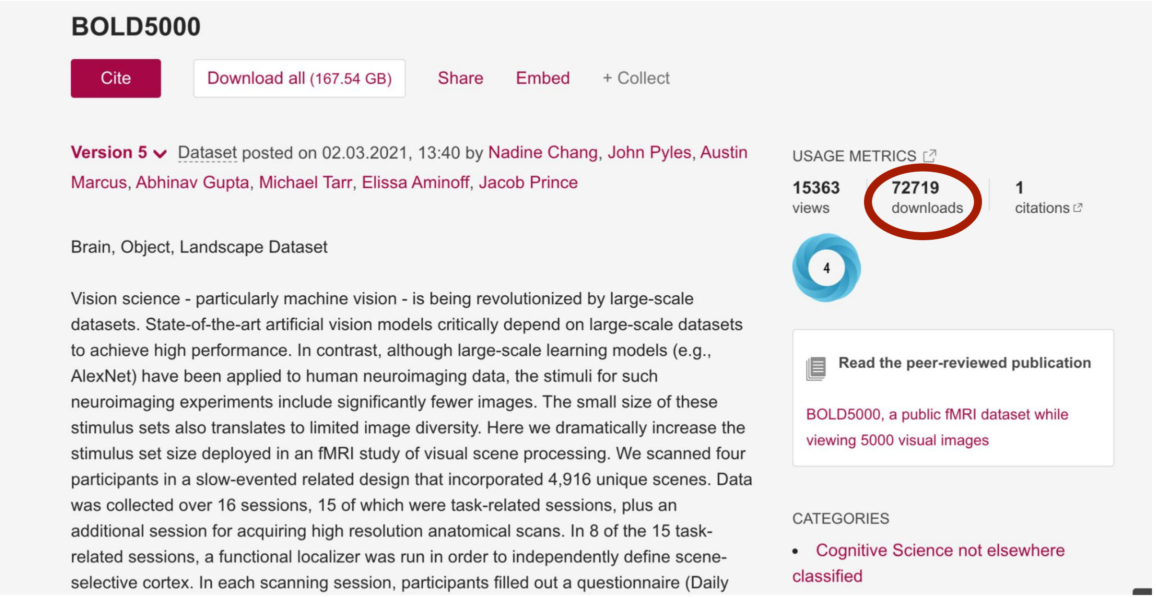
Search on KiltHub

Log in

Discover research from Carnegie Mellon University

https://kilthub.cmu.edu

Increase impact and data reuse



BOLD5000

Version 5 Dataset posted on 02.03.2021, 13:40 by Nadine Chang, John Pyles, Austin Marcus, Abhinav Gupta, Michael Tarr, Elissa Aminoff, Jacob Prince

Brain, Object, Landscape Dataset

Vision science - particularly machine vision - is being revolutionized by large-scale datasets. State-of-the-art artificial vision models critically depend on large-scale datasets to achieve high performance. In contrast, although large-scale learning models (e.g., AlexNet) have been applied to human neuroimaging data, the stimuli for such neuroimaging experiments include significantly fewer images. The small size of these stimulus sets also translates to limited image diversity. Here we dramatically increase the stimulus set size deployed in an fMRI study of visual scene processing. We scanned four participants in a slow-evented related design that incorporated 4,916 unique scenes. Data was collected over 16 sessions, 15 of which were task-related sessions, plus an additional session for acquiring high resolution anatomical scans. In 8 of the 15 task-related sessions, a functional localizer was run in order to independently define scene-selective cortex. In each scanning session, participants filled out a questionnaire (Daily

USAGE METRICS

15363	72719	1
views	downloads	citations

Read the peer-reviewed publication

BOLD5000, a public fMRI dataset while viewing 5000 visual images

CATEGORIES

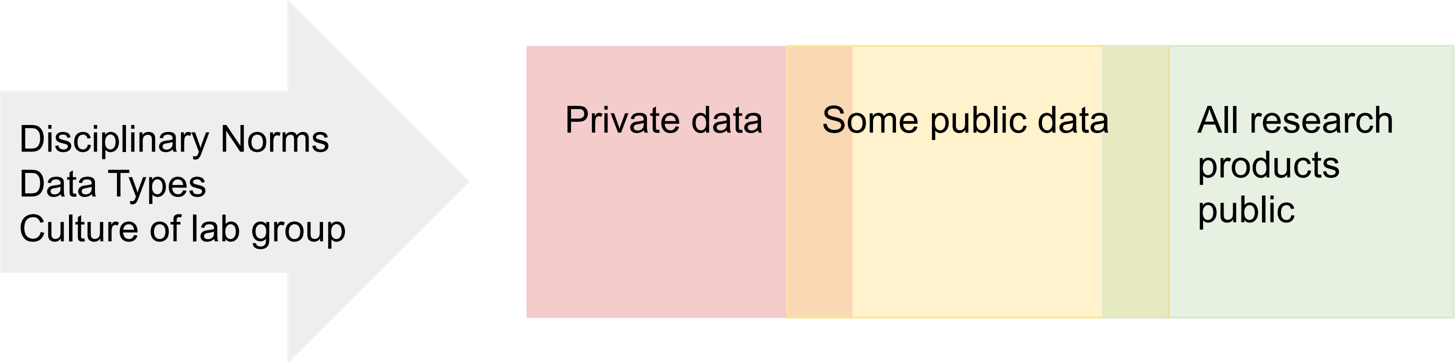
- Cognitive Science not elsewhere classified

https://kilthub.cmu.edu/articles/dataset/BOLD5000/6459449

Carnegie Mellon University

Program Outreach

- Presenting open science as a gradient of practices



Disciplinary Norms
Data Types
Culture of lab group

Private data

Some public data

All research products public

Individuals who are not familiar or interested in open science

Open Science practitioners, might be used to data sharing because of mandates

Open Science advocates and champions

Program Outreach

- Example of outreach for protocols.io

Private Protocols	Protocols shared with research group	Public protocols
• Improved documentation • Version Control • Reproducibility for your later self	• Reproducibility over time within a lab • Publishing complete methods	• Transparency • Discoverability • Importance for fixing the reproducibility crisis

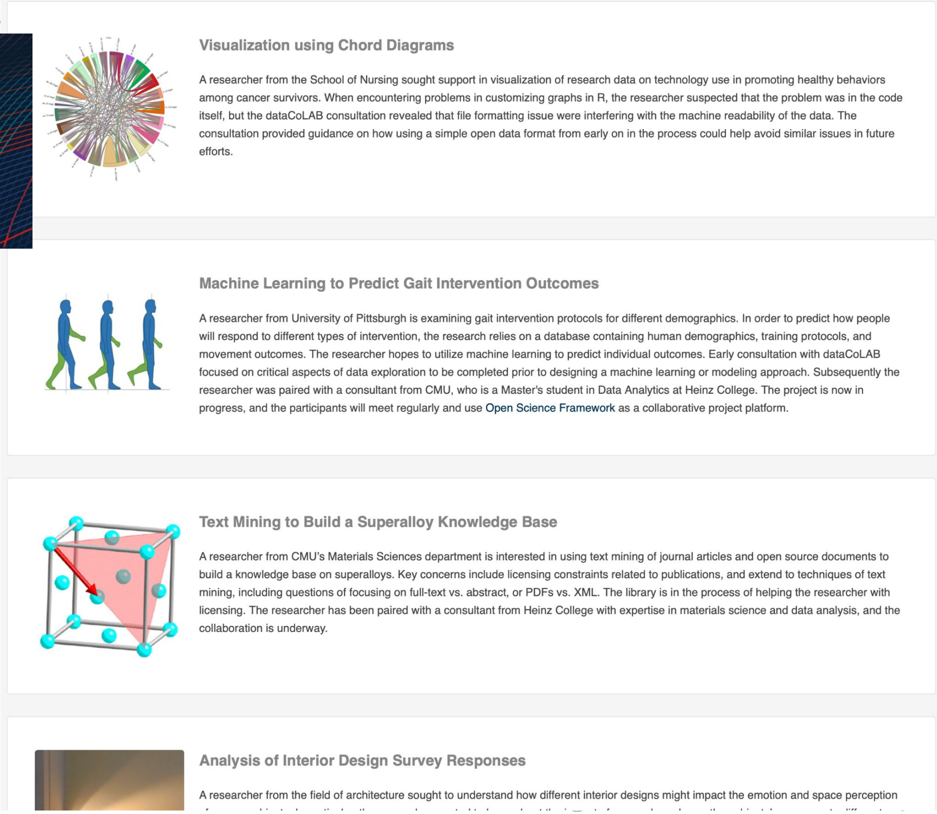
Creating internship-like opportunities with dataCoLAB

Data Collaborations Lab

Helping data producers and data scientists connect and collaborate.

https://cmu-lib.github.io/data-colab/

> *"This project started in a way because of COVID, … [dataCoLAB] gave me the confidence that, this is doable, and I don't have to do this all by myself."*
>
> *"I certainly did learn some new skills and used some of the work I've done only in theory on "real" datasets."*



Visualization using Chord Diagrams

A researcher from the School of Nursing sought support in visualization of research data on technology use in promoting healthy behaviors among cancer survivors. When encountering problems in customizing graphs in R, the researcher suspected that the problem was in the code itself, but the dataCoLAB consultation revealed that file formatting issue were interfering with the machine readability of the data. The consultation provided guidance on how using a simple open data format from early on in the process could help avoid similar issues in future efforts.

Machine Learning to Predict Gait Intervention Outcomes

A researcher from University of Pittsburgh is examining gait intervention protocols for different demographics. In order to predict how people will respond to different types of intervention, the research relies on a database containing human demographics, training protocols, and movement outcomes. The researcher hopes to utilize machine learning to predict individual outcomes. Early consultation with dataCoLAB focused on critical aspects of data exploration to be completed prior to designing a machine learning or modeling approach. Subsequently the researcher was paired with a consultant from CMU, who is a Master's student in Data Analytics at Heinz College. The project is now in progress, and the participants will meet regularly and use Open Science Framework as a collaborative project platform.

Text Mining to Build a Superalloy Knowledge Base

A researcher from CMU's Materials Sciences department is interested in using text mining of journal articles and open source documents to build a knowledge base on superalloys. Key concerns include licensing constraints related to publications, and extend to techniques of text mining, including questions of focusing on full-text vs. abstract, or PDFs vs. XML. The library is in the process of helping the researcher with licensing. The researcher has been paired with a consultant from Heinz College with expertise in materials science and data analysis, and the collaboration is underway.

Analysis of Interior Design Survey Responses

A researcher from the field of architecture sought to understand how different interior designs might impact the emotion and space perception

Increased need for data collaboration tools

Data Collaboration Tools Support Remote Research

'I began using LabArchives last fall, which has been intensely useful due to the Covid-19 pandemic. In March, when I prepared to work from home, I did not have to worry about taking home countless notebooks; I took my laptop home with me as usual.' - Sarah Werner, PhD candidate, Biological Sciences

Blog Post by Sarah Young: https://www.library.cmu.edu/about/news/2020-07/data-collaboration-tools-support-remote-research

1. Supporting open science at CMU and adapting to the pandemic
2. **Developing instruments for impact assessment**

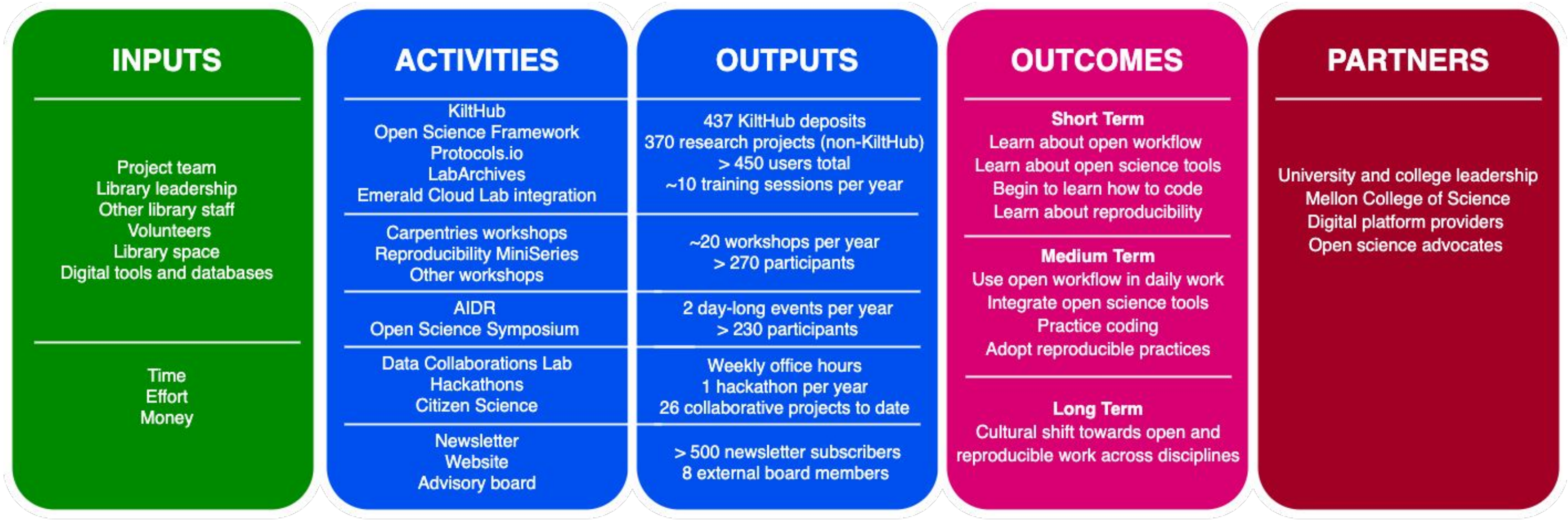
What impact are we making?

Who are our active users (and who are not)?

How does our program offerings benefit them?

How can we do more / better to support open science?

Developing a Logic Model



INPUTS

Project team
Library leadership
Other library staff
Volunteers
Library space
Digital tools and databases

Time
Effort
Money

ACTIVITIES

KiltHub
Open Science Framework
Protocols.io
LabArchives
Emerald Cloud Lab integration

Carpentries workshops
Reproducibility MiniSeries
Other workshops

AIDR
Open Science Symposium

Data Collaborations Lab
Hackathons
Citizen Science

Newsletter
Website
Advisory board

OUTPUTS

437 KiltHub deposits
370 research projects (non-KiltHub)
> 450 users total
~10 training sessions per year

~20 workshops per year
> 270 participants

2 day-long events per year
> 230 participants

Weekly office hours
1 hackathon per year
26 collaborative projects to date

> 500 newsletter subscribers
8 external board members

OUTCOMES

Short Term
Learn about open workflow
Learn about open science tools
Begin to learn how to code
Learn about reproducibility

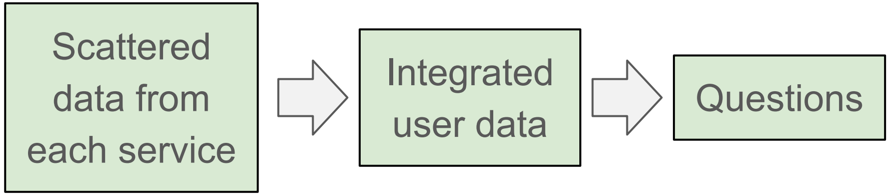
Medium Term
Use open workflow in daily work
Integrate open science tools
Practice coding
Adopt reproducible practices

Long Term
Cultural shift towards open and reproducible work across disciplines

PARTNERS

University and college leadership
Mellon College of Science
Digital platform providers
Open science advocates

Developing meaningful metrics



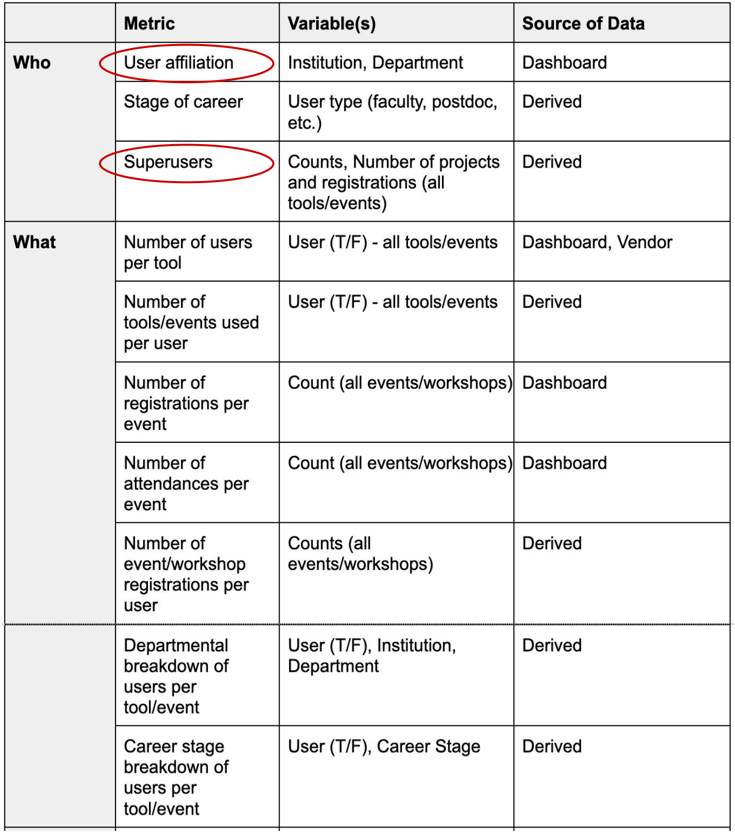
Scattered data from each service

Integrated user data

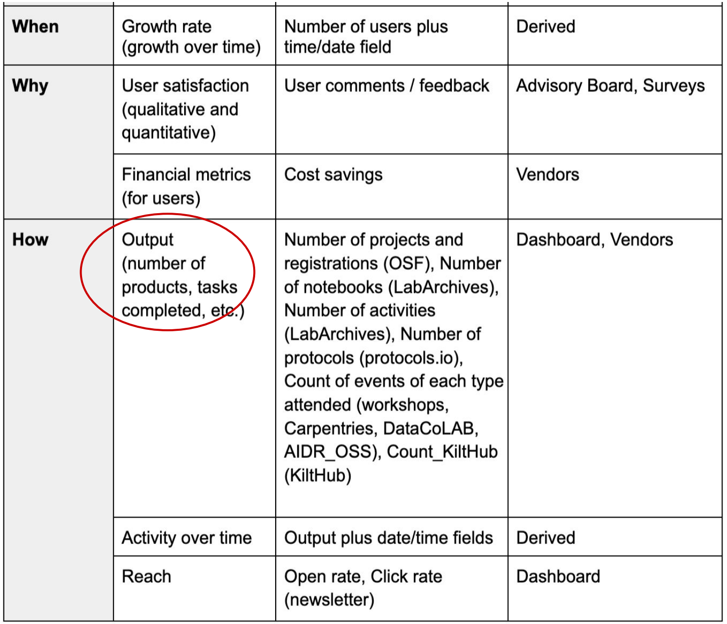
Questions

Who uses our tools and participates our activities?
Who are our top users?
Which **disciplines** are the most engaged?
How do people use our tools or activities?
Why do people use our tools or activities?
What **impact** are we making?

Developing meaningful metrics



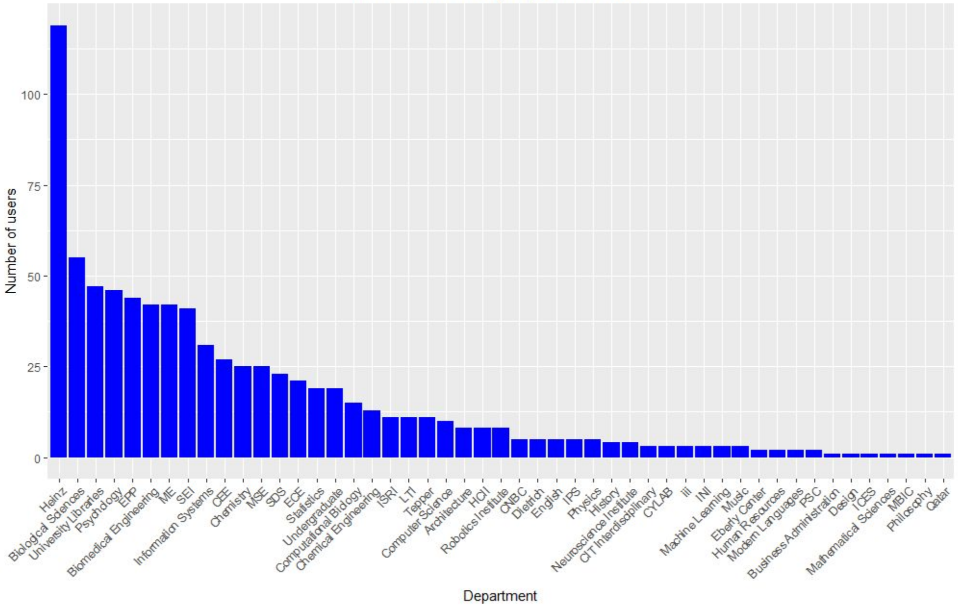
	Metric	Variable(s)	Source of Data
Who	User affiliation	Institution, Department	Dashboard
	Stage of career	User type (faculty, postdoc, etc.)	Derived
	Superusers	Counts, Number of projects and registrations (all tools/events)	Derived
What	Number of users per tool	User (T/F) - all tools/events	Dashboard, Vendor
	Number of tools/events used per user	User (T/F) - all tools/events	Derived
	Number of registrations per event	Count (all events/workshops)	Dashboard
	Number of attendances per event	Count (all events/workshops)	Dashboard
	Number of event/workshop registrations per user	Counts (all events/workshops)	Derived
	Departmental breakdown of users per tool/event	User (T/F), Institution, Department	Derived
	Career stage breakdown of users per tool/event	User (T/F), Career Stage	Derived



When	Growth rate (growth over time)	Number of users plus time/date field	Derived
Why	User satisfaction (qualitative and quantitative)	User comments / feedback	Advisory Board, Surveys
	Financial metrics (for users)	Cost savings	Vendors
How	Output (number of products, tasks completed, etc.)	Number of projects and registrations (OSF), Number of notebooks (LabArchives), Number of activities (LabArchives), Number of protocols (protocols.io), Count of events of each type attended (workshops, Carpentries, DataCoLAB, AIDR_OSS), Count_KiltHub (KiltHub)	Dashboard, Vendors
	Activity over time	Output plus date/time fields	Derived
	Reach	Open rate, Click rate (newsletter)	Dashboard

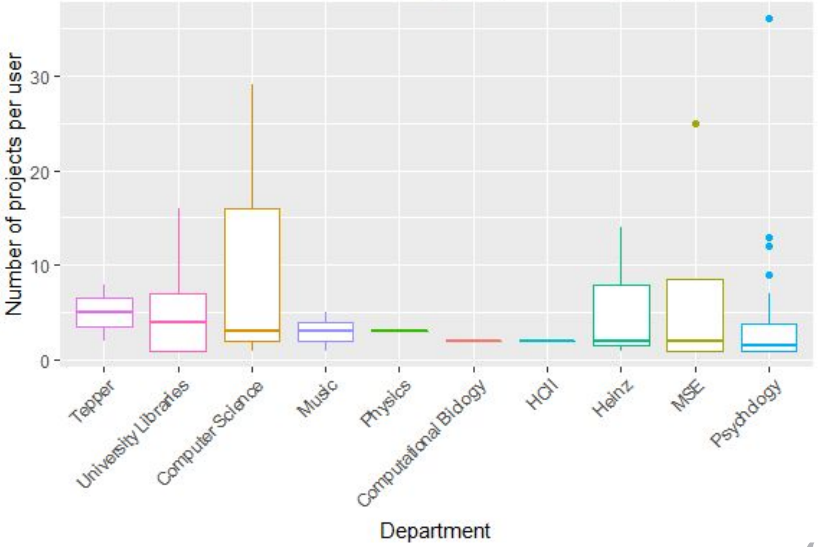
Carnegie Mellon University

Who: User affiliation (all data)


Number of users by department

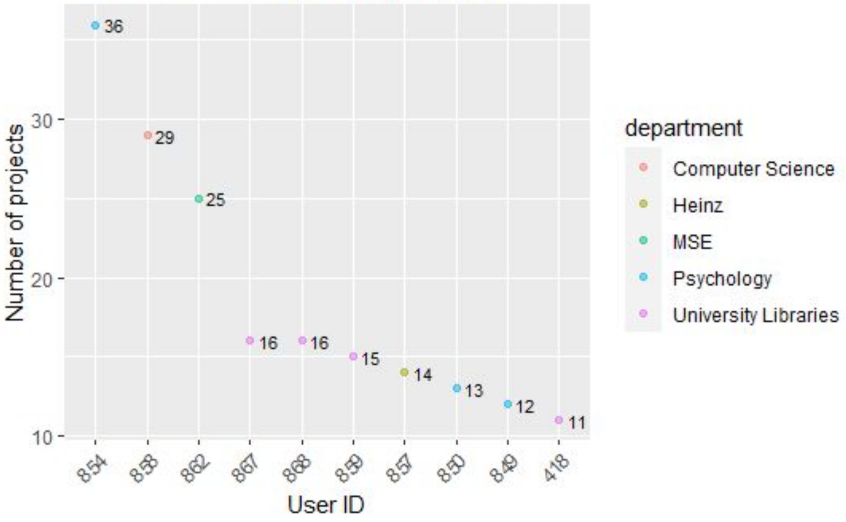
What / How: Departmental breakdown of projects

Distribution of number of KiltHub public items owned per user
(top 10 by median)



Who: Superusers

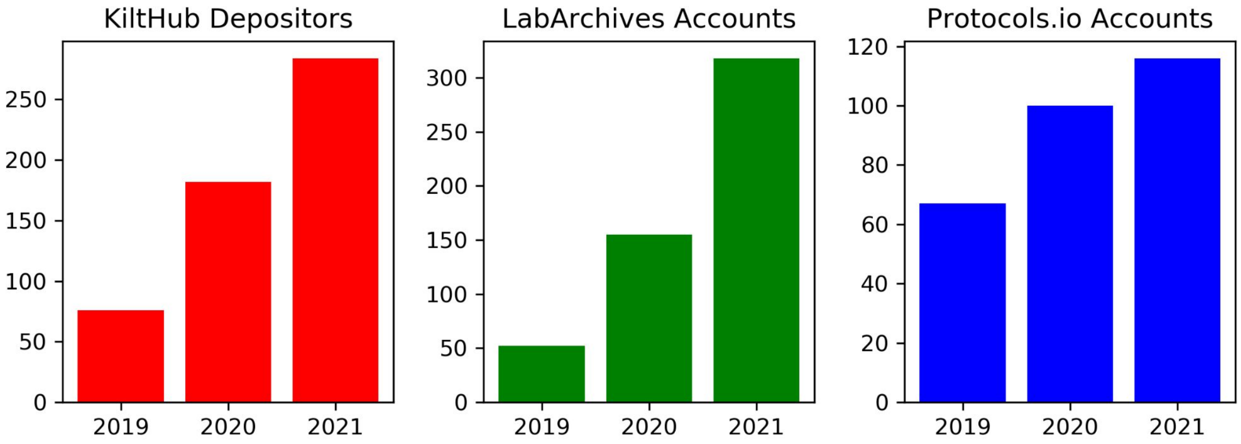
Number of KiltHub public items owned by superuser



Carnegie Mellon University

When: Growth over time

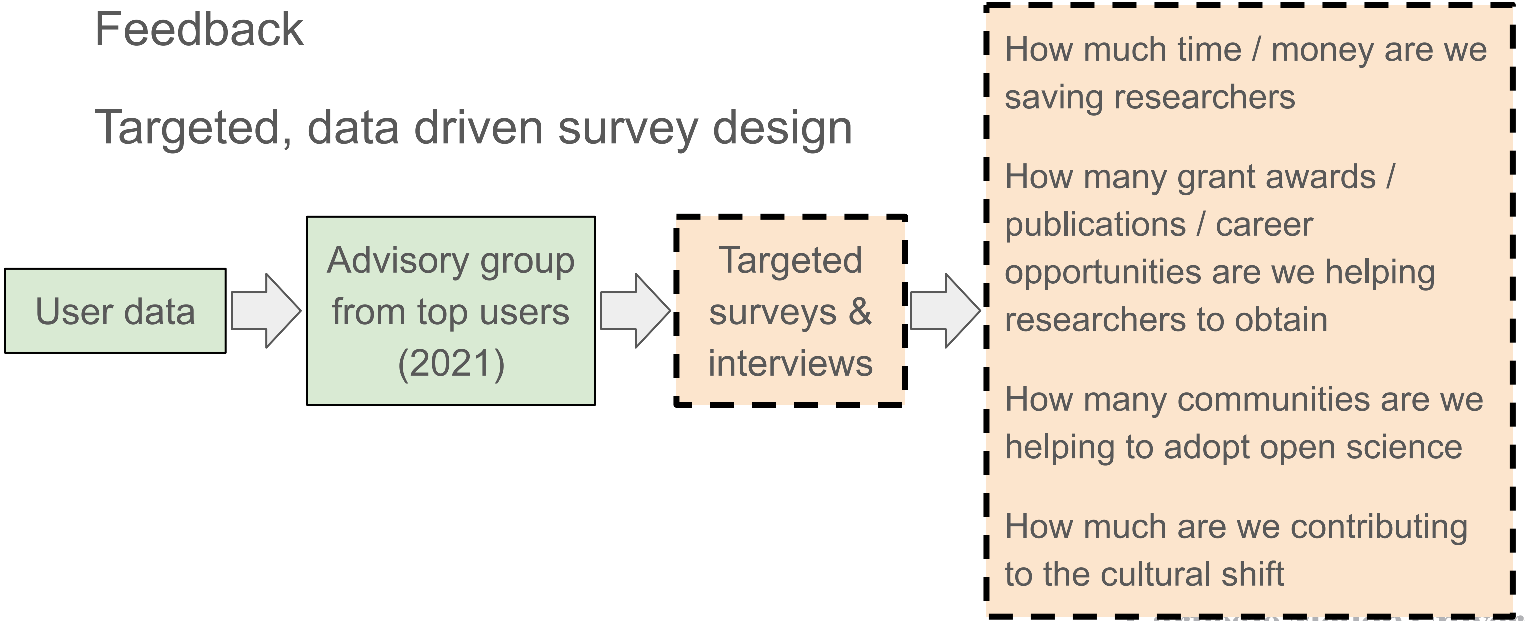
Cumulative user counts by year



Why: what do people get out of our services

Feedback

Targeted, data driven survey design



User data

Advisory group from top users (2021)

Targeted surveys & interviews

How much time / money are we saving researchers

How many grant awards / publications / career opportunities are we helping researchers to obtain

How many communities are we helping to adopt open science

How much are we contributing to the cultural shift

What impact are we making?

Acknowledgement

Ana Van Gulick

Sarah Young

Katie Behrman

Patrick Campbell

Chloe Woida

Hannah Gunderman

Emma Slayton

Neelam Bharti

Julie (Xiaoju) Chen

Matthew Lincoln

Let’s keep in touch!

Huajin Wang
huajinw@cmu.edu

Melanie Gainey
mgainey@andrew.cmu.edu

OSDC Program
✉ openscience@andrew.cmu.edu

#CMUOpenScience

https://www.library.cmu.edu/services/open-science