

Open Science Tools Within a Public Health Framework

By: Lance Picton, Karla Ajoy & Mehedi Hassan

Roots for Resilience (R4R) 2025

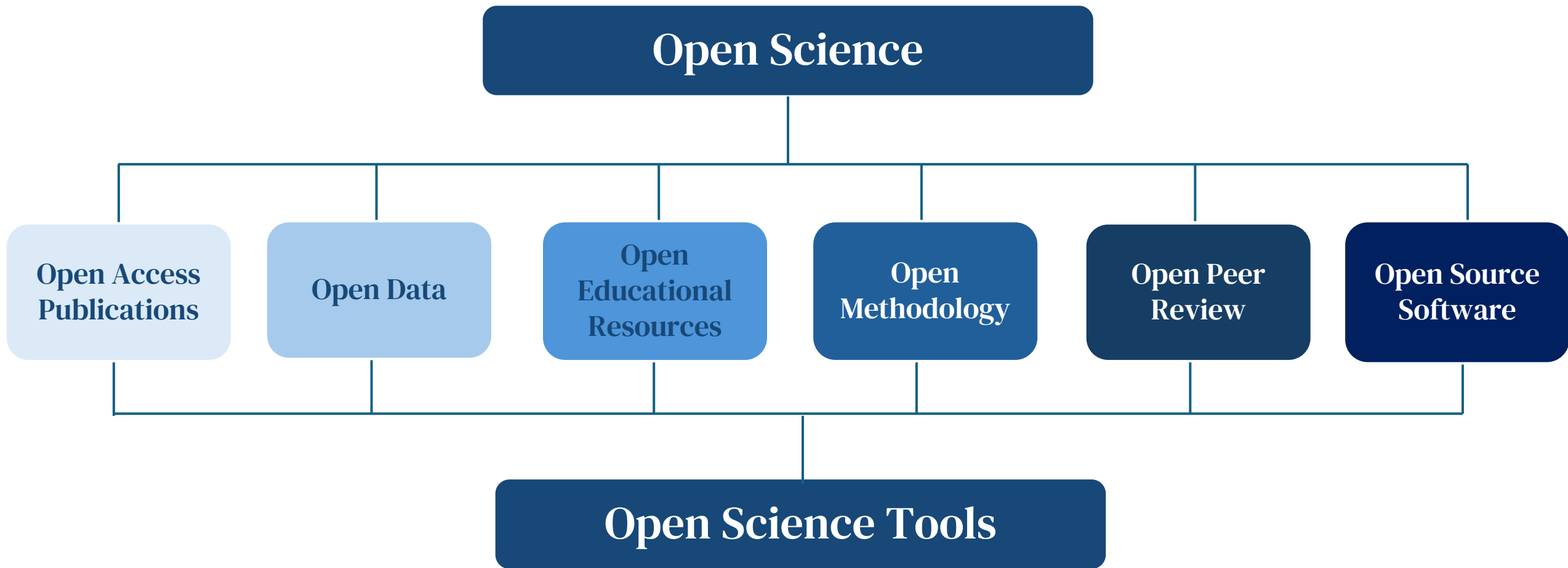


Open Science

“A collaborative and transparent approach to scientific research that emphasizes the accessibility, sharing, and reproducibility of data, methodologies, and findings to foster innovation and inclusivity”

-ChatGPT





Open Methodology

Platforms to collaboratively develop code and share with the open internet for reproducibility.



GitHub



GitLab



Open Source Softwares

•Linux operating system and shell



•Python



•R



•Git



•Conda



•Docker



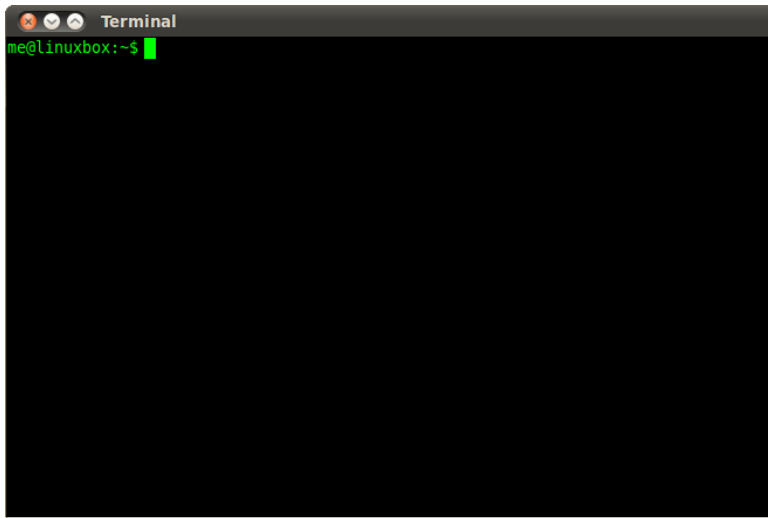
•Cyverse



•Pytorch



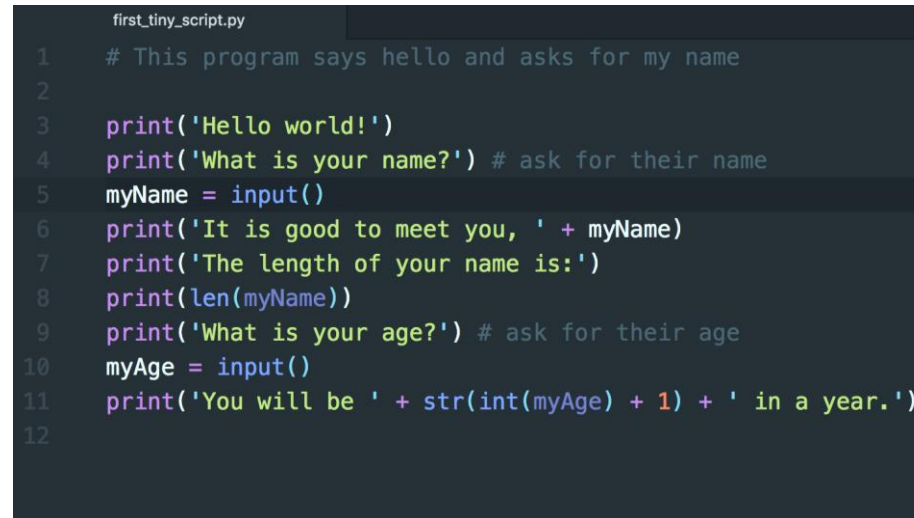
•Linux operating system and shell



```
Terminal
me@linuxbox:~$
```

- Open-source operating system
- Manages hardware, files and programs
- Users can customize with distributors (Ubuntu, Debian, etc.)

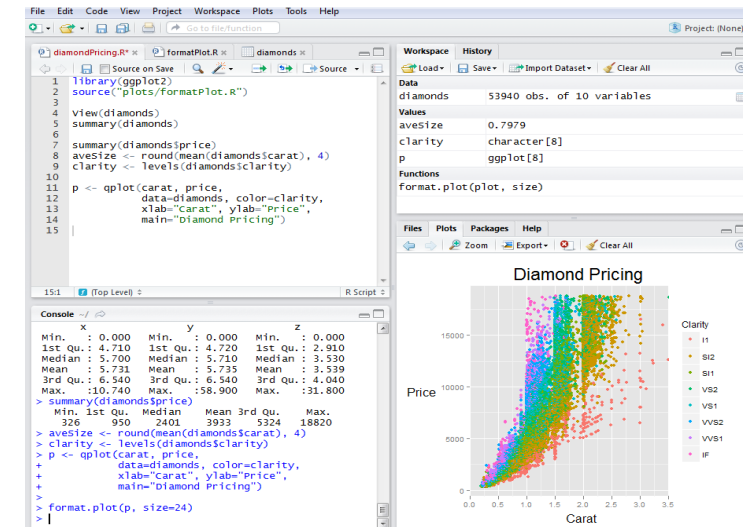
•Python



```
first_tiny_script.py
1 # This program says hello and asks for my name
2
3 print('Hello world!')
4 print('What is your name?') # ask for their name
5 myName = input()
6 print('It is good to meet you, ' + myName)
7 print('The length of your name is:')
8 print(len(myName))
9 print('What is your age?') # ask for their age
10 myAge = input()
11 print('You will be ' + str(int(myAge) + 1) + ' in a year.')
12
```

- Programming language
- Data science, machine learning, automation and web development
- Contains many libraries (Pandas, NumPy, TensorFlow, etc.)

•R



- Programming language
- Statistics, data analysis and visualization
- Heavy statistical modelling



•Git

```
MINGW64/c/Users/singh/Desktop/newRepo
nothing to commit, working tree clean

singh@DESKTOP-PGVSHMF MINGW64 ~/Desktop/newRepo (master)
$ git add .

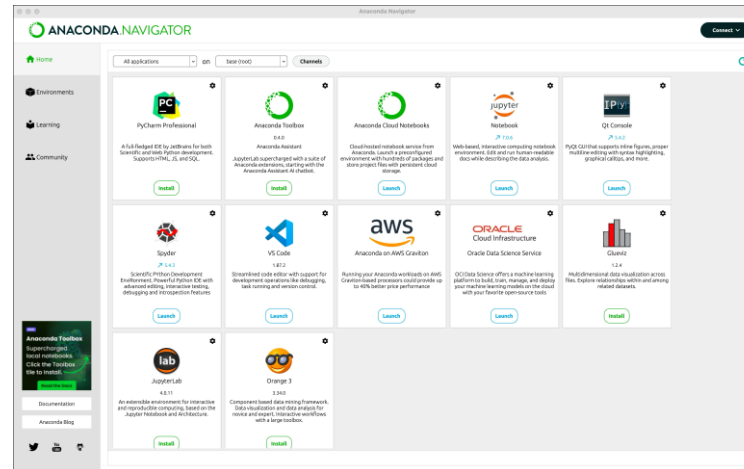
singh@DESKTOP-PGVSHMF MINGW64 ~/Desktop/newRepo (master)
$ git commit -m "second commit"
[master 9e7f7d0] second commit
1 file changed, 1 insertion(+)
create mode 100644 abc/jhvjhb.txt

singh@DESKTOP-PGVSHMF MINGW64 ~/Desktop/newRepo (master)
$ git push origin master
Enumerating objects: 5, done.
Counting objects: 100% (5/5), done.
Delta compression using up to 8 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (4/4), 326 bytes | 65.00 KiB/s, done.
Total 4 (delta 0), reused 0 (delta 0)
To https://github.com/taran910/NewRepo.git
7a5d54b..9e7f7d0 master -> master

singh@DESKTOP-PGVSHMF MINGW64 ~/Desktop/newRepo (master)
$
```

- Version control to track changes in coding files
- Allows branching, merging and rollback
- Employed with GitHub and GitLab to share and store repositories

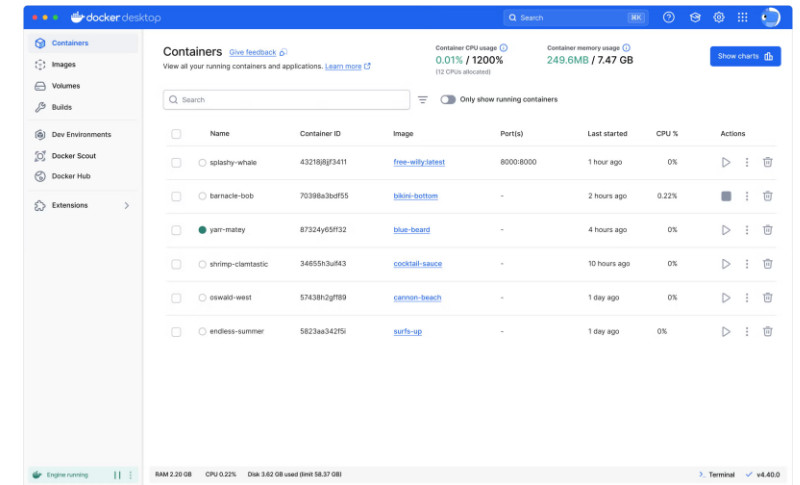
•Conda



- Environment manager that helps install software and have project dependencies isolated
- Supports different programming languages
- Usually used through Anaconda or Miniconda

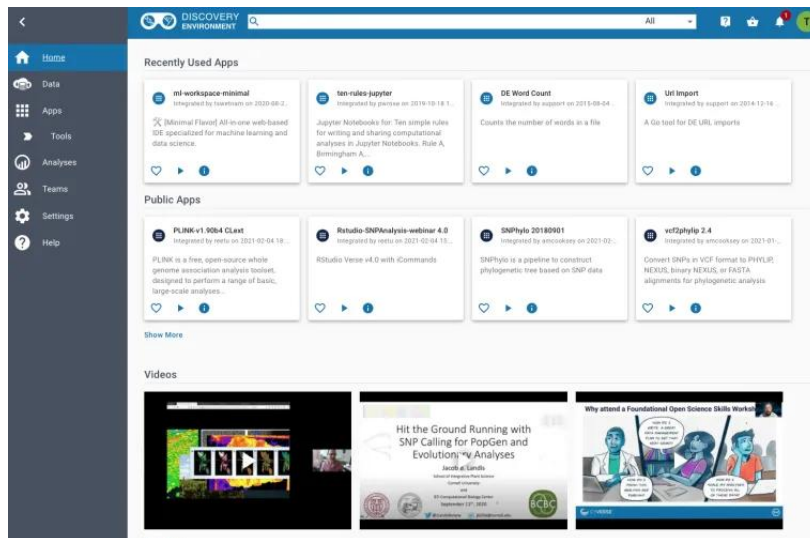


•Docker



- Platform that packages applications and dependencies into ‘containers’
- Ensure software runs the same way
- Used in reproducible research

•Cyverse



- Cloud-based research platform that allows data storage, HPC and analysis tools
- Supports big biological datasets, workflows and reproducible research

•Pytorch

```
import torch
from functorch import functionalize
from torch.fx.experimental.proxy_tensor import make_fx

def f(x):
    y = x.clone()
    y.add_(1)
    return y

functionalized_f = functionalize(f)
x = torch.ones(4)
print(torch.allclose(f(x), functionalized_f(x)))

# Print an FX graph of the "functionalized" version of f
fx_g = make_fx(functionalized_f)(x)
print(fx_g.code)
```

- Deep learning framework to create and train neural networks
- Used for AI prototyping and Natural Language Processing (NLP)



Artificial Intelligence

•Ollama



•Bard



•ChatGPT



•Geminis



Pretrained models

•Hugging Face



HUGGING FACE

Pretrained model



| **Hugging Face: Models, Datasets, and Spaces**



Datasets

Thousands of curated datasets for ML tasks

Saves time in data collection and pre-processing

Some include a dataset viewer for interacting with data



Models

Pre-trained ML models

Can be used to perform diverse ML tasks across including text, audio, and image

Supports dozens of libraries, including Transformers



Spaces

Built, host, and share ML apps in a single, organized workspace

Thousands of ML-powered demo apps

Hosted with built-in Python frameworks (Streamlit or Gradio) or JS and HTML (static).

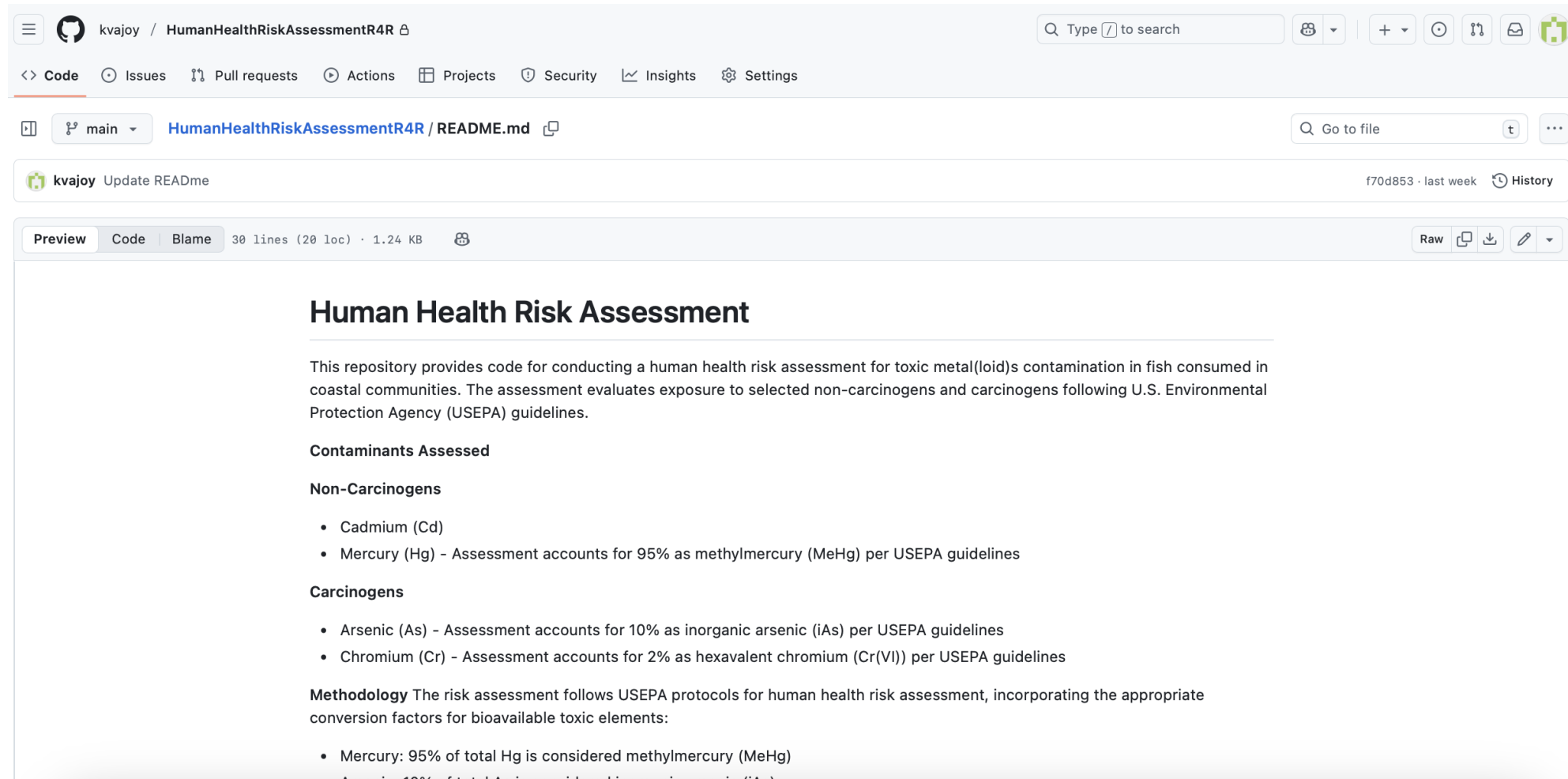


Open Methodology

```
1  ````{r data, include=FALSE}
2  library(readxl)
3  fish = read_excel("/Users/karla/Downloads/Master's/Data analysis MK/R4R HHRA.xlsx", sheet = "Fish") #load your source of document with the data
4  ````
5
6
7
8  C (mg/kg) is the concentration of metals in wet weight
9  IR (kg/day) is the ingestion rate
10 EF (day/year) is the exposure frequency (days e.g. 365 days for those
11 who eat shellfish seven times a week and 52 days for those who eat
12 shellfish once a week in a year)
13 ED (years) is the duration of exposure
14 BW (kg) is the body weight
15 AT ((day/year) x year) is the averaging time of non-carcinogens using the life
16 expectancy
17
18
19 ````{r HHRA concentration normal, include=FALSE}
20 monte_carlo_HRA <- function(
21   mean, sd,                #for concentration (C) - normal
22   min, max,                #min/max for concentration
23   n,
24   IR_meanlog, IR_sdlog, IR_min, IR_max, #lognormal IR with limits
25   EF_meanlog, EF_sdlog, EF_min, EF_max, #lognormal EF with limits
26   ED_min, ED_max,          #uniform ED
27   BW_mean, BW_sd,          #normal BW with range check
```



Open Methodology



The screenshot shows a GitHub repository page for 'HumanHealthRiskAssessmentR4R' by user 'kvajoy'. The repository is on the 'main' branch, and the selected file is 'README.md'. The file has 30 lines of code, 20 locations, and is 1.24 KB in size. The README content is as follows:

Human Health Risk Assessment

This repository provides code for conducting a human health risk assessment for toxic metal(loid)s contamination in fish consumed in coastal communities. The assessment evaluates exposure to selected non-carcinogens and carcinogens following U.S. Environmental Protection Agency (USEPA) guidelines.

Contaminants Assessed

Non-Carcinogens

- Cadmium (Cd)
- Mercury (Hg) - Assessment accounts for 95% as methylmercury (MeHg) per USEPA guidelines

Carcinogens

- Arsenic (As) - Assessment accounts for 10% as inorganic arsenic (iAs) per USEPA guidelines
- Chromium (Cr) - Assessment accounts for 2% as hexavalent chromium (Cr(VI)) per USEPA guidelines

Methodology

The risk assessment follows USEPA protocols for human health risk assessment, incorporating the appropriate conversion factors for bioavailable toxic elements:

- Mercury: 95% of total Hg is considered methylmercury (MeHg)
- Arsenic: 10% of total As is considered inorganic arsenic (iAs)



"Those who fail to plan, plan to fail"

DMP?

"A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed."





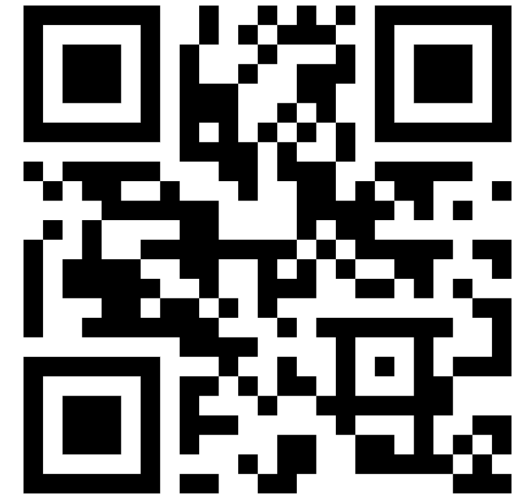
University of Arizona Libraries

Data Management and Sharing Plan Tool (DMPTool)

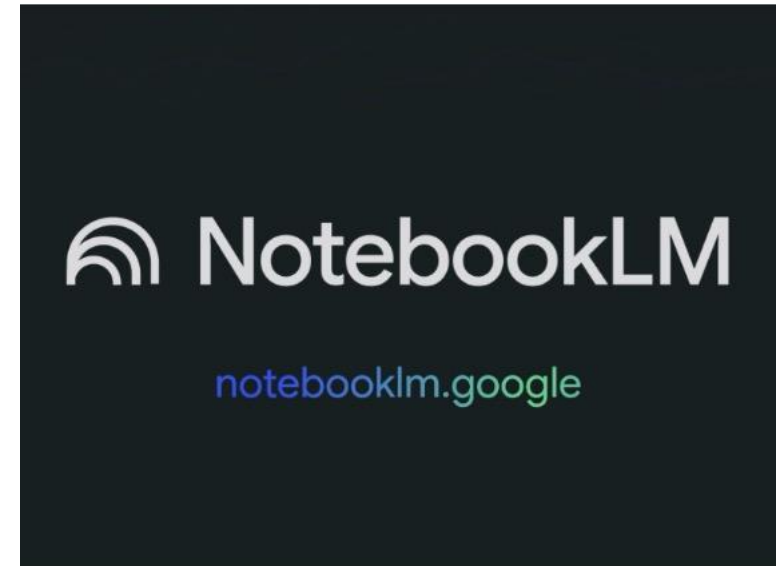
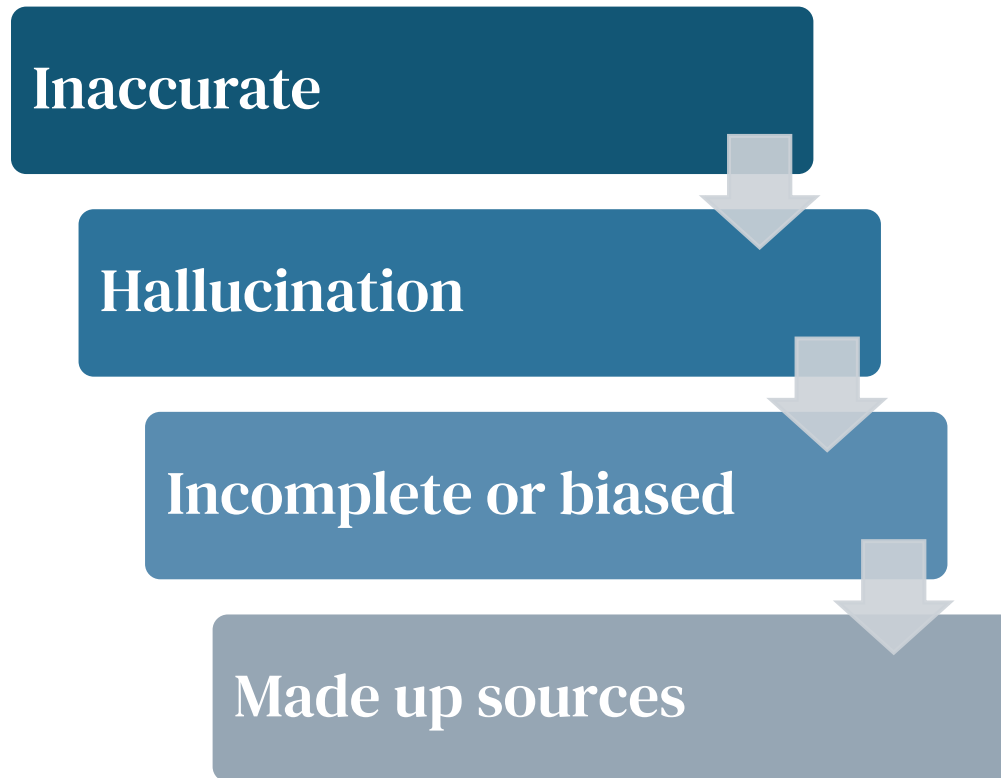
RESUME

DETAILS ▼

- ☰ What is the DMPTool?
- ☰ Logging In as a University of Arizona User
- ☰ Navigating the DMPTool Dashboard
- ☰ Creating a Plan
- ☰ Adding Project Details to Your Plan
- ☰ Adding and Managing Contributors
- ☰ Adding and Managing Collaborators
- ☰ Writing Your Plan
- ☰ Identifying Research Outputs
- ☰ Downloading Your Final Grant Proposal



Literature review & AI



Paper & Author Networks



Select None Select All

• GEE

☒ Gorelick ☐ Moore 2017
3302

Google Earth Engine: Planetary-scale geospatial analysis for everyone
Remote Sensing of Environment

☐ Wu 2020
29

Geemap: A Python package for interactive mapping with Google Earth Engine
J. Open Source Softw.

☐ Wu ☐ Lang 2019
51

Integrating LIDAR data and multi-temporal aerial imagery to map wetland inundation dynamics using Google Earth Engine.
Remote Sensing of Environment

☐ Pekel ☐ Belward 2016
1591

High-resolution mapping of global surface water and its long-term changes
Nature

☐ Hansen ☐ Townshend 2013
5702

High-Resolution Global Maps of 21st-Century Forest Cover Change
Science

☐ Kumar ☐ Mutanga 2018
170

Google Earth Engine Applications Since Inception: Usage, Trends,

+ Add Papers

Connections
Click to Show

☒ Selected paper

☐ Noel Gorelick ☐ Rebecca Moore

Google Earth Engine: Planetary-scale geospatial analysis for everyone
Remote Sensing of Environment 2017
3392

[PDF](#)

Abstract Google Earth Engine is a cloud-based platform for planetary-scale geospatial analysis that brings Google's massive computational capabilities to bear on a variety of high-impact societal issues including deforestation, drought, disaster, disease, food security, water management, climate monitoring and environmental protection. It is unique in the field as an integrated platform designed to empower not only traditional remote sensing scientists, but also a much wider audience that lacks the technical capacity needed to utilize traditional supercomputers or large-scale commodity cloud computing resources.

1 Selected Paper

Remove from:
• GEE

Add to Other Collection

EXPLORE PAPERS

[Similar Work](#) 998

[All References](#) 30

[All Citations](#) 3383

EXPLORE PEOPLE

[These Authors](#) 8

[Suggested Authors](#) 8

EXPLORE OTHER CONTENT

[Linked Content](#) 1

EXPORT PAPERS

BibTeX RIS CSV

PUBLIC COLLECTION ☒

SHAREABLE LINK [Copy](#)

COLLABORATORS [Edit](#)

EMAIL UPDATES ☒

Filter Relevance

☐ Abstracts ☒ Comments

Select All

Pekel ☐ Belward 2016
1591

High-resolution mapping of global surface water and its long-term changes
Nature

Hansen ☐ Townshend 2013
5792

High-Resolution Global Maps of 21st-Century Forest Cover Change
Science

Drusch ☐ Bargellini 2012
1641

Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services
Remote Sensing of Environment

Wulder ☐ Woodcock 2012
752

Opening the archive: How free data has enabled the science and monitoring promise of Landsat
Remote Sensing of Environment

Lewis ☐ Wang 2017
147

The Australian Geoscience Data Cube - foundations and lessons learned
Remote Sensing of Environment

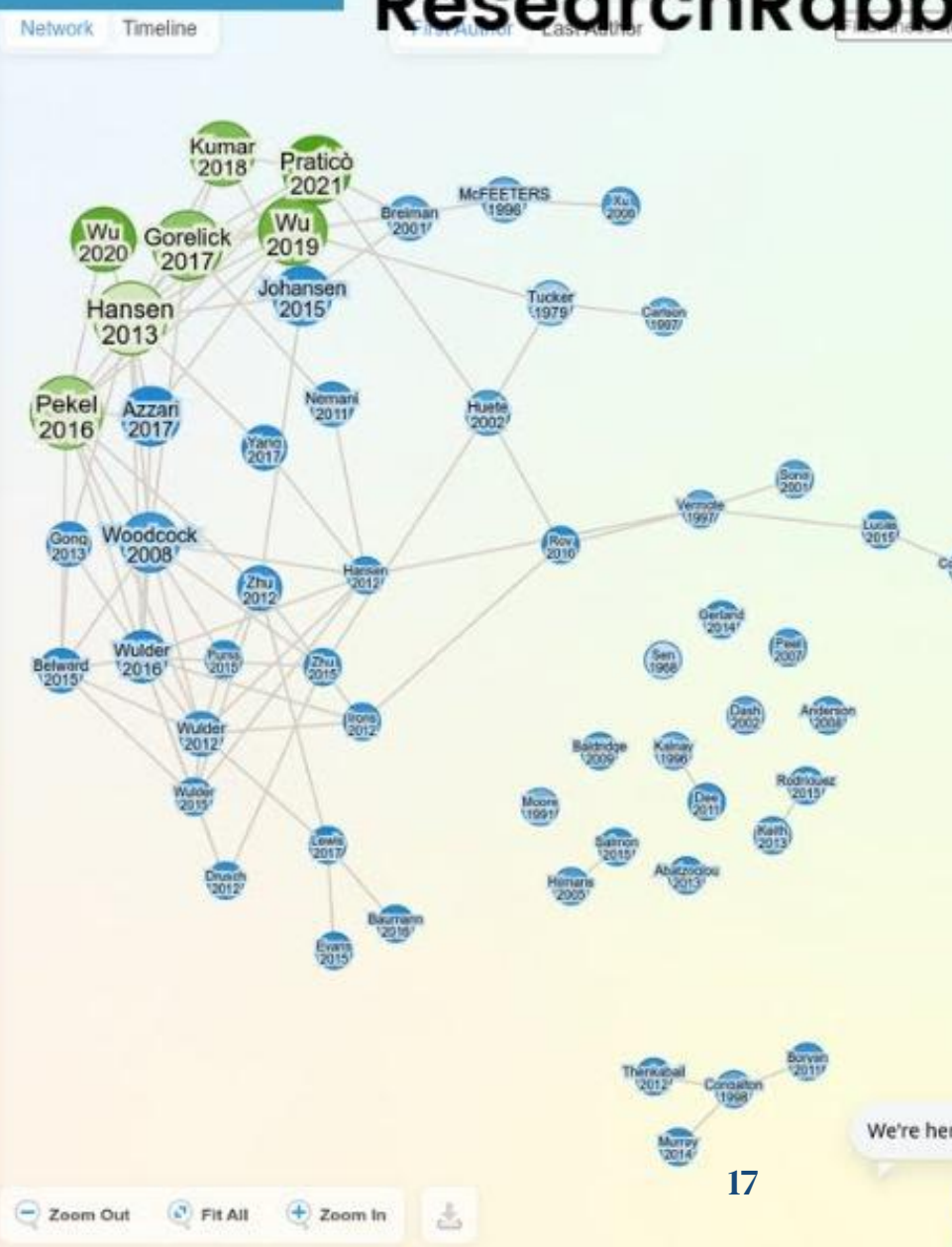
Breiman 2001
5953

Random Forests

Hansen ☐ Loveland 2012
624

A review of large area monitoring of land cover change using Landsat data

Connections
Click to Hide





Sources

+ Add sources

Try Deep Research for an in-depth report and new sources!

Search the web for new sources

Web Fast Research

Select all sources

PDF

Essentials_of_Epidemiology_in_Public...

✓

PDF

Essentials_of_Epidemiology_in_Public...

✓

PDF

Essentials_of_Epidemiology_in_Public...

✓

PDF

Essentials_of_Epidemiology_in_Public...

✓

PDF

Essentials_of_Epidemiology_in_Public...

✓

PDF

Essentials of Epidemiology in Public...

✓

Chat

Untitled notebook

6 sources

These academic materials provide a comprehensive overview of the fundamental principles and methodologies required for epidemiological study in public health. The texts explain the essential metrics for measuring disease, contrasting the occurrence of new cases, or **incidence**, with the frequency of existing cases, known as **prevalence**, while also defining different population types and calculation methods such as **rates and proportions**. These measures form the basis of **descriptive epidemiology**, which analyzes disease distribution according to the characteristics of **person, place, and time** to generate hypotheses and guide resource allocation. Furthermore, the sources explore the intricate process of **causal inference**, detailing necessary attributes like **time order and association** and employing systematic frameworks such as **Hill's guidelines** to establish genuine causes rather than mere risk factors. Finally, a critical section is dedicated to the **ethical oversight of research**, emphasizing the

Start typing...

What are the foundational differences between measuring disease incidence, prevalence, and comparing resultant rates?

How do epidemiologists use these metrics to establish causal inference?

6 sources

Studio

Audio Overview

Video Overview

Mind Map

Reports

Flashcards

Quiz

Infographic

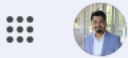
Slide Deck

Studio output will be saved here.

After adding sources, click to add Audio Overview, Study Guide, Mind Map, and more!

Add note

Notebook M can be inaccurate; please double check its responses.



+

PDF

PDF

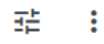
PDF

PDF

PDF

PDF

Chat ⓘ



Today • 9:07 PM

hill's guideline

The guidelines proposed by Sir Austin Bradford Hill are a set of nine viewpoints suggested in his much-cited 1965 publication to help determine if an observed association is causal 1 2 . Hill introduced these guidelines in the context of debates over the causal nature of the association between smoking and lung cancer 3 .

These are often referred to as "causal criteria," although Hill and modern epidemiologists more appropriately term them **guidelines** 2 Hill emphasized that they were meant to be imperfect guides to aid epidemiologists in forming a judgment about causality, rather than rigid criteria 2 .

The nine guidelines suggested by Sir Austin Bradford Hill for assessing causation are 2 5 :

1. **Strength of association** 5 : A large association is considered more likely to be causal than a small 5 primarily because strong associations are unlikely to be accounted for entirely by alternative explanations such as bias and confounding 6 7 . However, Hill noted

Start typing... 6 sources →

Define and explain Rothman's sufficient-component cause model.

List the three major principles of the Belmont Report.

Which Hill principle is the only causality requirement?



Chat ⓘ



The guidelines proposed by Sir Austin Bradford Hill are a set of nine viewpoints suggested to determine if an observed association is causal ¹ ² . Hill introduced these guidelines in the association between smoking and lung cancer ³ .

These are often referred to as "causal criteria," although Hill and modern epidemiologists Hill emphasized that they were meant to be imperfect guides to aid epidemiologists in formulating criteria ² .

The nine guidelines suggested by Sir Austin Bradford Hill for assessing causation are ² ⁵ :

1. **Strength of association** ⁵ : A large association is considered more likely to be causal than a small one primarily because strong associations are unlikely to be accounted for entirely by alternative explanations such as bias and confounding ⁶ ⁷ . However, Hill noted

TABLE 15-3 Hill's Guidelines for Assessing Causation

■ Strength of association ■ Consistency ■ Specificity
■ Temporality ■ Biological gradient ■ Plausibility ■
Coherence ■ Experiment ■ Analogy

Hill's Guidelines for Assessing Causation 399

consistency

Hill asserted that associations are more likely to be causal if they are observed repeatedly "by different persons, in different places, circumstances and times."

This is because replication increases our confidence that the association is not due to an error or fallacy. For example, one of the reasons that epidemiologists believe that smoking causes lung cancer is that this association has been observed in so many different studies

conducted by different investigators in different

hill's guideline

Start typing...

6 sources →

Define and explain Rothman's sufficient-component cause model.

List the three major principles of the Belmont Report.

Which Hill principle is the only causality requirement?

Documentation On Markdown

Table of contents

- 1 ADMINISTRATIVE PROCEDURES
- 2 ANALYTICAL PROCEDURES
- 3 SPECTROSCOPY
- 4 MINERALOGY
- 5 PHYSICAL PROPERTIES
- 6 LAB TOOLS & EXPERIMENTS
- 7 PAPERS AND REFERENCES

Distribute to items -> In item type for Chemicals is Quantity tax exempt, but for any other supplies Quantity Taxable.

- At the bottom click CALCULATE
- SUBMIT

[Back to top](#)

2 ANALYTICAL PROCEDURES

ALEC Sample volume required for different analyses: [ALEC Analytes](#)

[ICP-MS \(main cations\) & Arsenic Speciation](#) [Ion Chromatography \(Anions\)](#) [Iron Speciation](#)

[Ferrozine Method for Iron Speciation](#) [TOC \(Total Organic Carbon\)](#) [pH for mine tailings](#)

Analysis	Volume needed
ICP-MS / ICP-AES / Arsenic Speciation	3 mL
Total Organic Carbon, Total Carbon and Total Nitrogen	17 mL
Ion Chromatography	1 mL
HPLC ICP-MS	300 μ L (<i>micro Liters</i>)
HPLC electrospray ionization (ESI) MS	1 mL
Ferrozine method	ALEC does not run this, Fan/Fan method
TOTAL VOLUME NEEDED PER SAMPLING POINT:	25 mL

Kind of Tubes to use for this sample preparation:

- 15 mL or 50 mL VWR Centrifuge Tube
- Clear PP - Metal Free
- Sterilize - Flat Cap
- They need to be acid-washed with 2% HNO₃ before use
- PLUS point: 15 mL fit inside the sample collector of Dr. Markus Tuller

Metal Free Tubes


3 SPECTROSCOPY

XAS FTIR

[The Orange Book](#) [Our Synchrotron Data](#) [Sample prep formats](#) [LARCH Python package](#)

[SIXPACK SOFTWARE](#) [ATHENA SOFTWARE](#) [Databases](#)

Download the Orange Book [here](#)



Distribute to Items -> In Item Type for Chemicals is Quantity exempt, but for any other supplies Quantity Taxable.

- At the bottom click CALCULATE
- SUBMIT

[Back to top](#)

2 ANALYTICAL PROCEDURES

5

ALEC Sample volume required for different analyses:

[ALEC Analytes](#)

[ICP-MS \(main cations\) & Arsenic Speciation](#)

[Ion Chromatography \(Anions\)](#)

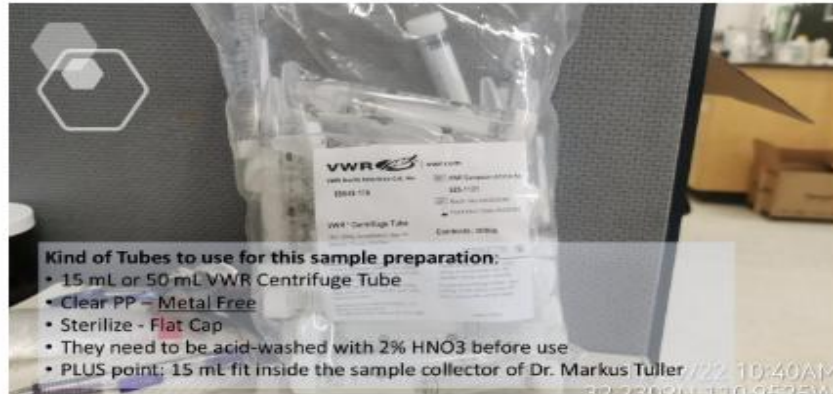
[Iron Speciation](#)

[Ferrozine Method for Iron Speciation](#)

[TOC \(Total Organic Carbon\)](#)

[pH for mine tailings](#)

Analysis	Volume needed
ICP-MS / ICP-AES / Arsenic Speciation	3 mL
Total Organic Carbon, Total Carbon and Total Nitrogen	17 mL
Ion Chromatography	1 mL
HPLC ICP-MS	300 μ L (<i>micro Liters</i>)
HPLC electrospray ionization (ESI) MS	1 mL
Ferrozine method	ALEC does not run this, Fan/Fan method
TOTAL VOLUME NEEDED PER SAMPLING POINT:	25 mL



Metal Free Tubes

3 SPECTROSCOPY

[XAS](#) [FTIR](#)

[The Orange Book](#)

[Our Synchrotron Data](#)

[Sample prep formats](#)

[LARCH Python package](#)

[SIXPACK SOFTWARE](#)

[ATHENA SOFTWARE](#)

[Databases](#)

Download the Orange Book [here](#)



MK Docks Hosted On GitHub


You can even develop your portfolio!

Mehedi Hasan

Search

Home

Table of contents
Mehedi Hasan
My Journey





Mehedi Hasan

Doctoral Student & Graduate Research Associate
Environmental Health Sciences
The University of Arizona, USA

My Journey

Mehedi Hasan's journey in public health began in the vibrant yet challenging landscapes of Bangladesh, where he witnessed firsthand the pressing need for sustainable solutions. After earning his Bachelor's and Master's degrees in Public Health & Informatics at Jahangirnagar University, he spent five impactful years at the International Centre for Diarrhoeal Disease Research Bangladesh (icddr,b) working with marginalized communities to address critical WASH challenges. Now, as a second-year PhD student & Graduate Research Associate in UofA's Community, Environment, and Policy department, Mehedi is developing innovative health projects and mentoring future leaders, driven by a lifelong commitment to equity and global change. **More in About...**



Education

Explore my academic journey, from public health to environmental health sciences.

Making Research Reproducible

"Reproducing the result of a computation means running the same software on the same input data and obtaining the same results." Rougier et al. 2016

Software Dependency Hell

- Sometimes, it can be nearly impossible to get your computing environment correct enough to run someone else's code.
- This can be caused by incorrect software versions of the packages you are using or their dependencies.
- Don't Dispair!

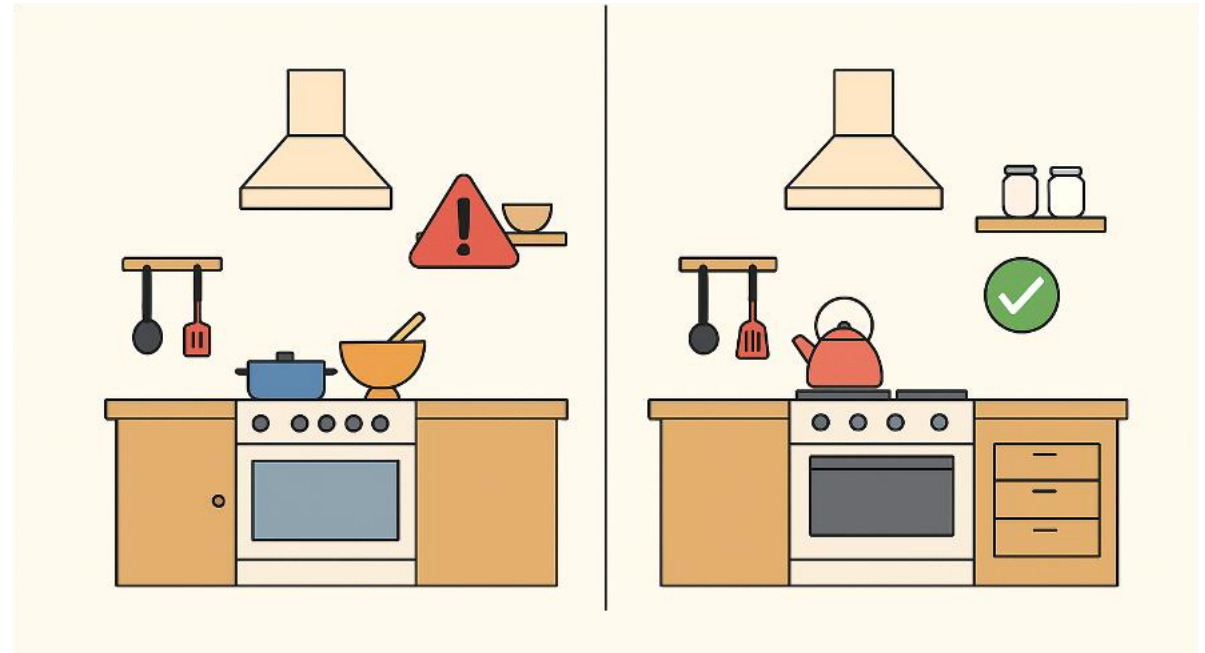
Software Environments & Docker Container

There are solutions to avoid software dependency hell and ensure reproducibility from one computer to another



What is a Software Environment?

- An environment manager allows you to create software installation directories that are isolated from your computer's PATH
- You can create unique environments and install specific software version to run specific scripts



Imagine your research runs inside a 'kitchen' of tools (software), recipes (code), and ingredients (data).

If the kitchen setup changes, the recipe fails!

What is Docker container?

- A useful analogy is to think of software containers as shipping containers. It allows us move cargo (software) around the world in standard way. The shipping container can be offloading and executed anywhere, as long the destination has a shipping port (i.e., Docker)



Software environment = kitchen
Docker = container keeping the kitchen consistent

Containers For Reproducible Science

- **Environment Consistency:**
 - Containers encapsulate the software environment, ensuring that the same versions of software, libraries, and dependencies are used every time,
 - reducing the "it works on my machine" problem.
- **Ease of Sharing:**
 - Containers can be easily shared with other researchers, allowing them to replicate the exact software environment used in a study.
- **Platform Independence:**
 - Containers can run on different operating systems and cloud platforms, allowing for consistency across different hardware and infrastructure.
- **Version Control:**
 - Containers can be versioned, making it easy to keep track of changes in the software environment over time.
- **Scalability:**
 - Containers can be easily scaled and deployed on cloud infrastructure, allowing for reproducible science at scale.
- **Isolation:**
 - Containers isolate the software environment from the host system, reducing the risk of conflicts with other software and ensuring a clean and controlled environment.

R4R- Barriers to Success



Research Group/PI



Institution/University



Government

UA OPEN ACCESS POLICY

Open by Default

A university-wide commitment to
global knowledge exchange.

*"We support open practices—including
open data and open source software."*

- Universal Access for all scholars
- Transparent & reproducible code
- Equity in global research

The Applicability Gap

70%

International Students

Supportive Ecosystem U of A:

- HPC Access
- Mentorship
- Open Science Policies



RESEARCH & PARTNERSHIPS
Data Science Institute

The Limits of Influence

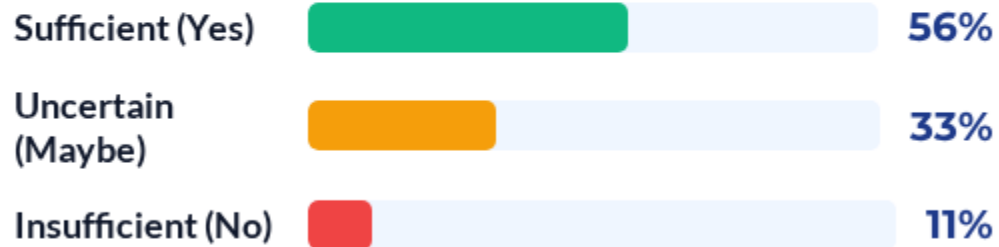
However, this influence is geographic.

- University's institutional power
- Infrastructure support

Infrastructure Bottlenecks

When asked about the sufficiency of home computing infrastructure (HPC, cloud access, reliable internet), the cohort response was split:

Sufficiency of Home Computing Infrastructure

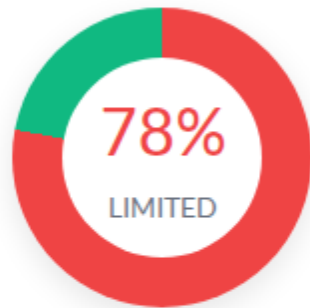


Key Challenge: Instability

- **44% of researchers** expressed uncertainty
- Divide in accessibility and applicability

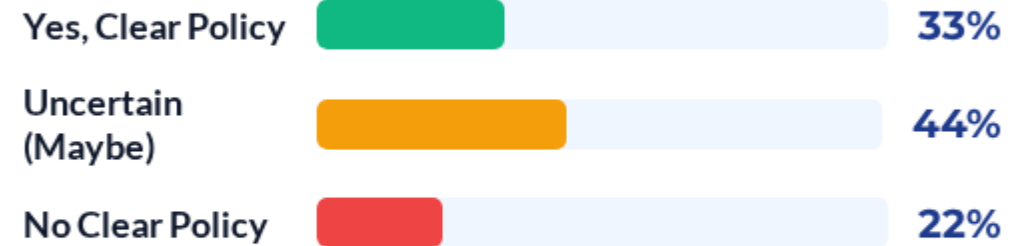
Policy & Legal Hurdles

Anticipated Data Localization Limits



The vast majority anticipate their research will be ****significantly limited****

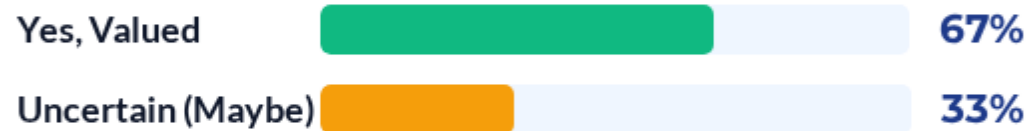
Clear Open Science Policies



Only 1 in 3 institutions currently have clear policies supporting

Cultural & Acceptance Friction

Senior Researchers Reward Open-Source



****33% uncertainty**** highlights a lack of explicit, formalized incentives.

Likelihood of Tool Acceptance

Acceptance by Country: **56% Definitely Yes**

Acceptance by Institution : **56% Definitely Yes**

The Challenge:

Over half are confident in acceptance, but this confidence is undermined by the hard data on insufficient infrastructure and limiting data laws.

Perceived Openness: High Intent

The perceived willingness to use R4R/FOSS tools is overwhelmingly high, demonstrating strong motivation at every level:

4 in 5

Home Country

Average willingness to utilize R4R tools.

15%

Institution/University

Average perceived disinterestedness from
the local institution.

86%

Research Cohort/PI

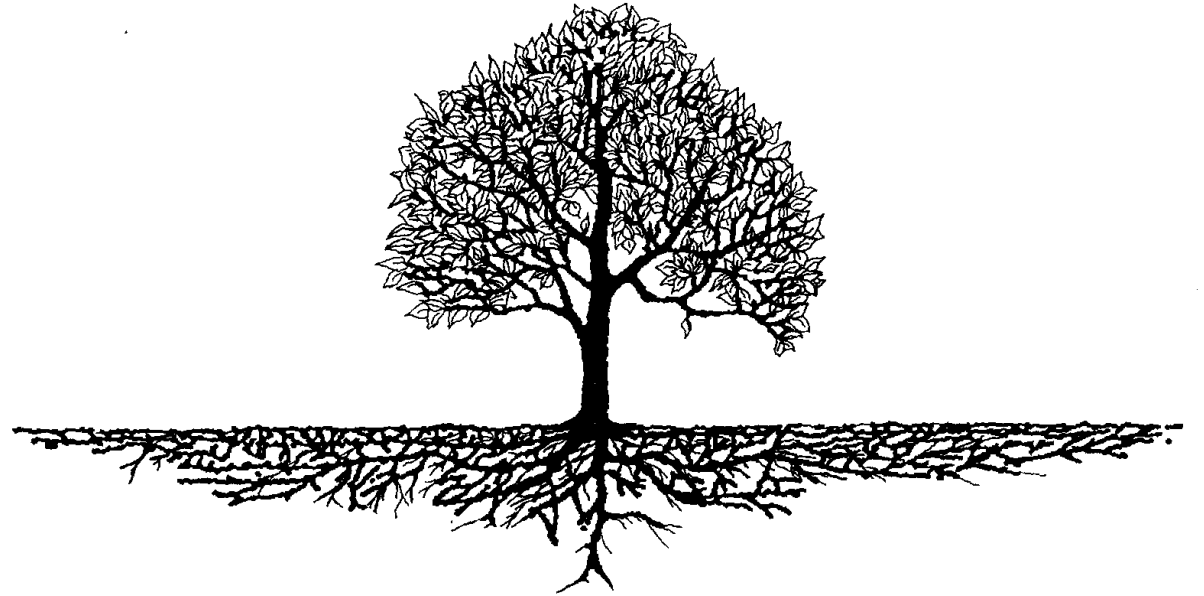
Highest score reflects strong support from
peers and mentors.

CONCLUSION: The problem is not one of motivation, but of **enabling environment****.**

Building True Resilience

Thriving in non-ideal conditions.

- ✓ **Adapt:** Create low-resource, containerized tool versions.
- ✓ **Advocate:** Push for local(international) Open Science policies.
- ✓ **Connect:** Build regional alumni support networks.





Mel & Enid Zuckerman
College of Public Health



Lance Picton
jackpicton@arizona.edu



Karla Ajoy
kvajoy@arizona.edu



Mehedi Hasan
mehedihasan@arizona.edu



RESEARCH & PARTNERSHIPS
Data Science Institute



**Arizona Institute
for Resilience**