Cold Spring Harbor Laboratory

Researcher Perspectives – NIH Generalist Repository Ecosystem Initiative (GREI) Workshop

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
@JasonWilliamsNY

Generalist Repository Ecosystem Initiative (GREI) Workshop, Online
January 2023

DOWNLOAD SLIDES

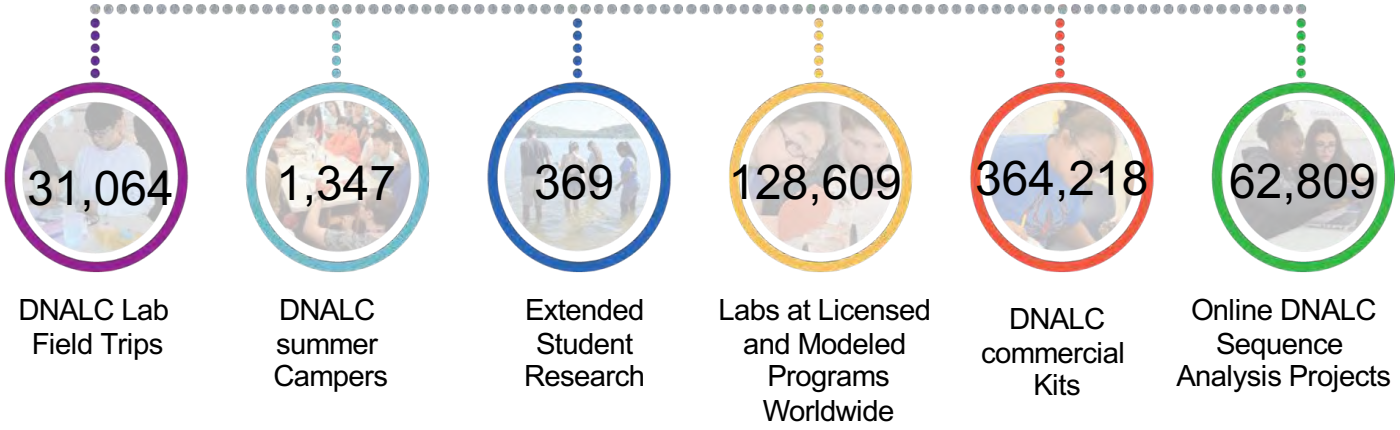
bit.ly/grei-jan-2023


Cold Spring Harbor Laboratory

“Prepare students and families to thrive in the gene age…”

CSHL DNA Learning Center

588,416 Annual Exposures

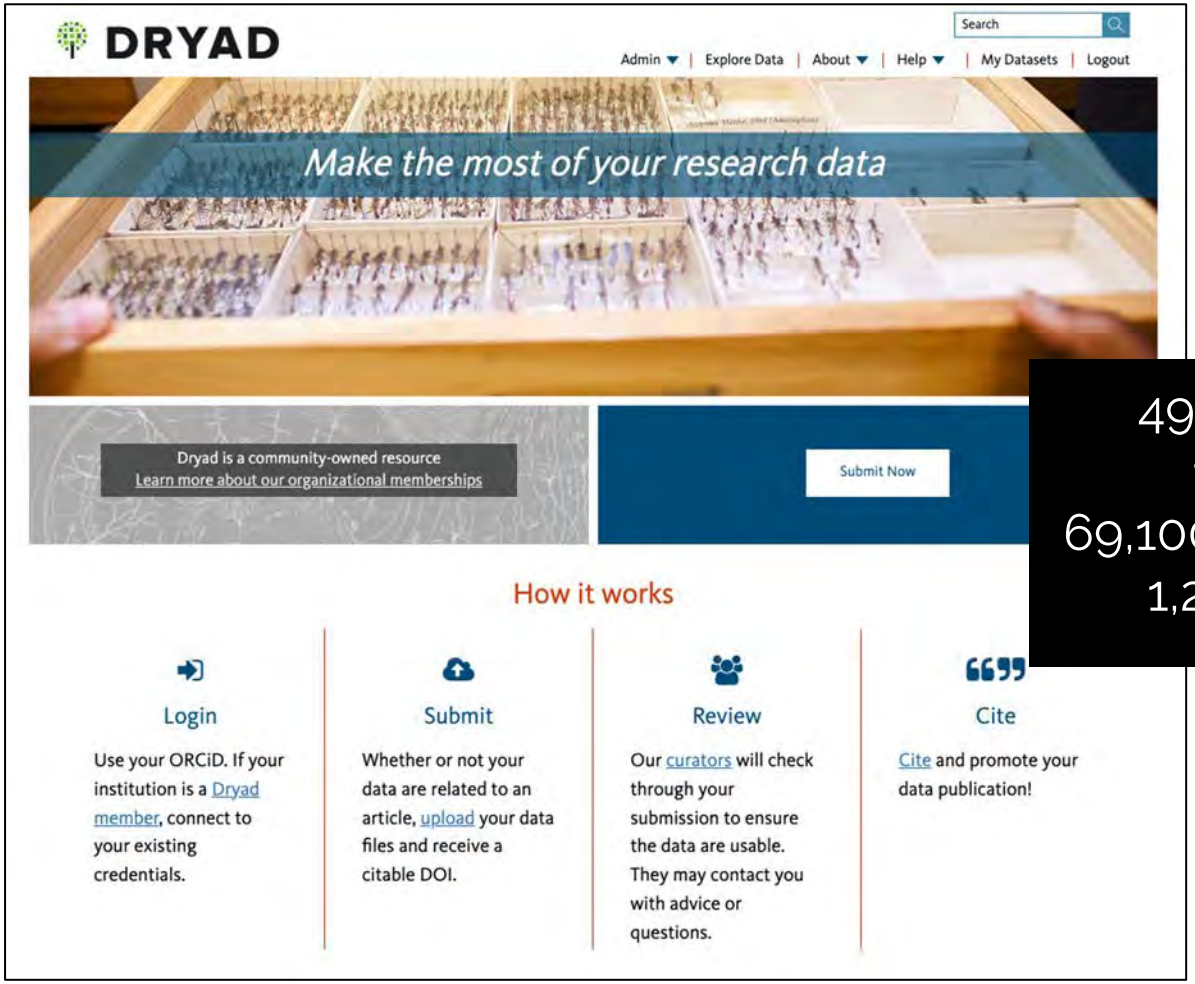


An **open data publishing platform & community** committed to the **open availability and routine re-use** of all research data

AN OPEN DATA PUBLISHING PLATFORM & COMMUNITY

Vision: **All research data is openly available and routinely re-used**

Mission: **Enable and promote** the re-use of research data

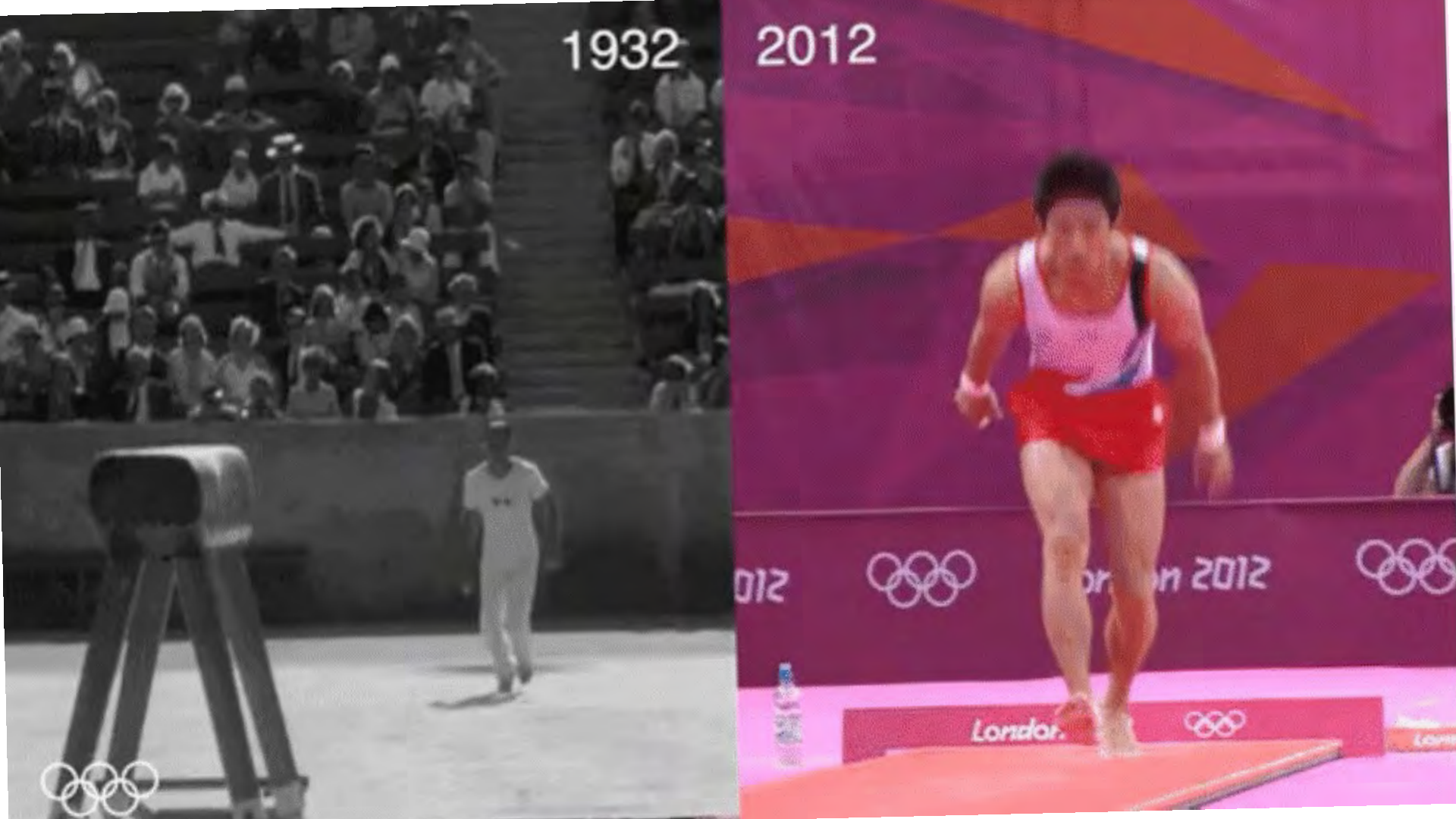


49,600+ data publications
193,300+ researchers
69,100+ international institutions
1,270+ academic journals

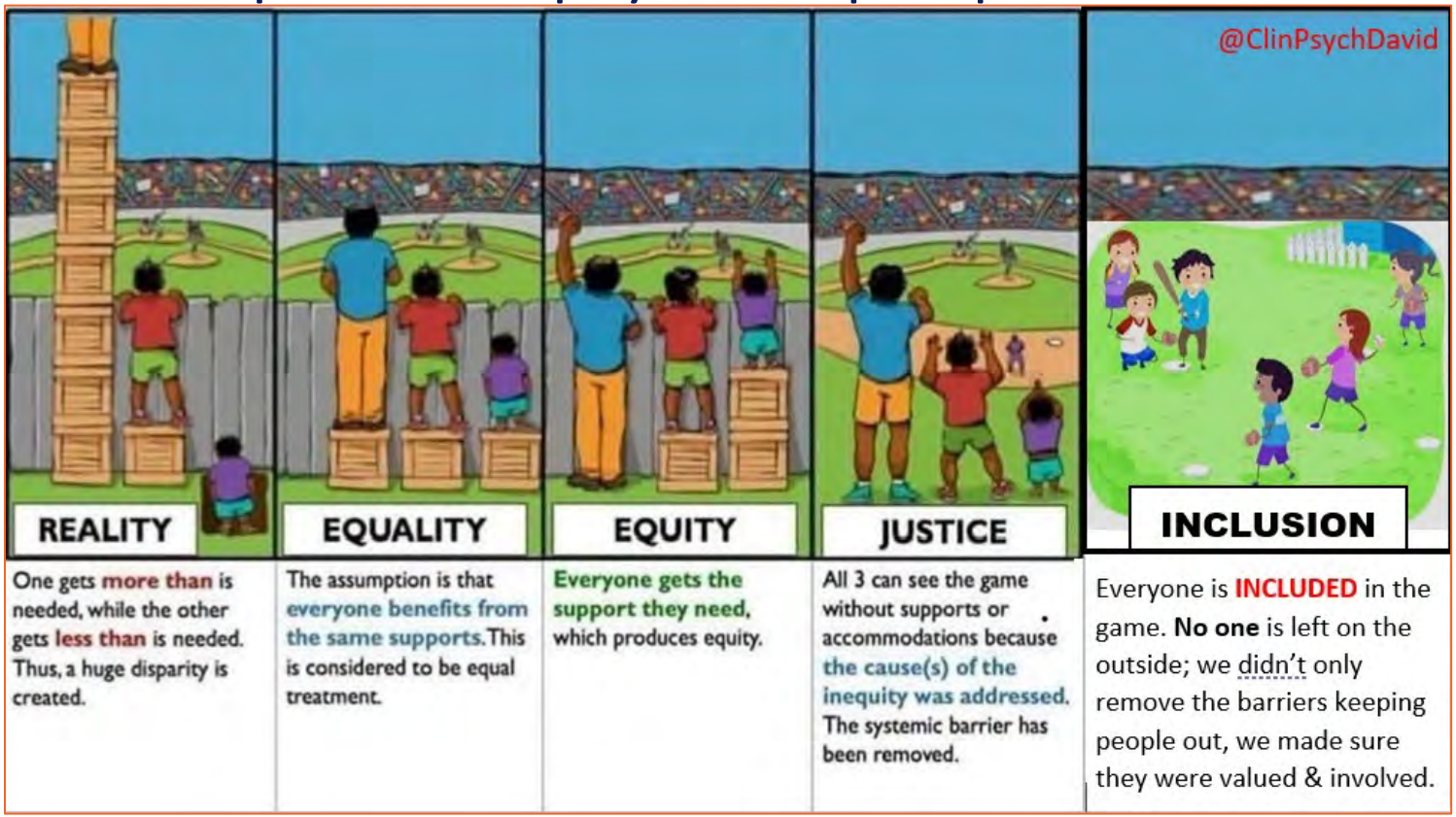
I went to school for…

I need to know this for my research…

```
import urllib2
eutils = 'http://www.ncbi.nlm.nih.gov/entrez/eutils/'
efetch = 'efetch.fcgi?'
s = eutils + efetch

targets = ['J04243','M60064']
idString = 'id=' + ','.join(targets)
s += idString + '&db=nucleotide&rettype=fasta'
fileObject = urllib2.urlopen(s)
data = fileObject.read().strip()

entries = data.split('\n\n')
title, sequence = entries[0].split('\n', 1)
print title.split(' ',1)[0]

# prints:
# >gi|154102|gb|J04243.1|STYHEMAPRF
```


1932
2012

Gaps multiply and perpetuate

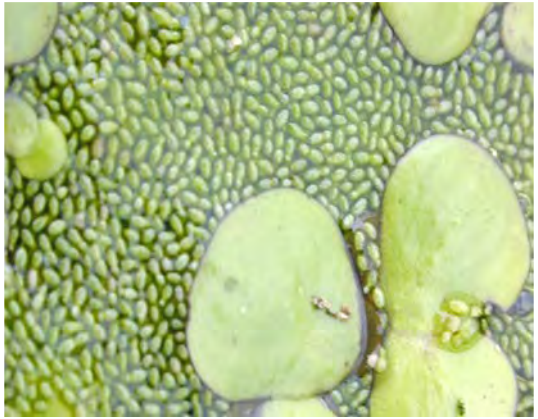


Credit: Saskatoon Health Service; Revised version - David Murphy

Sequence-a-genome camp

2021

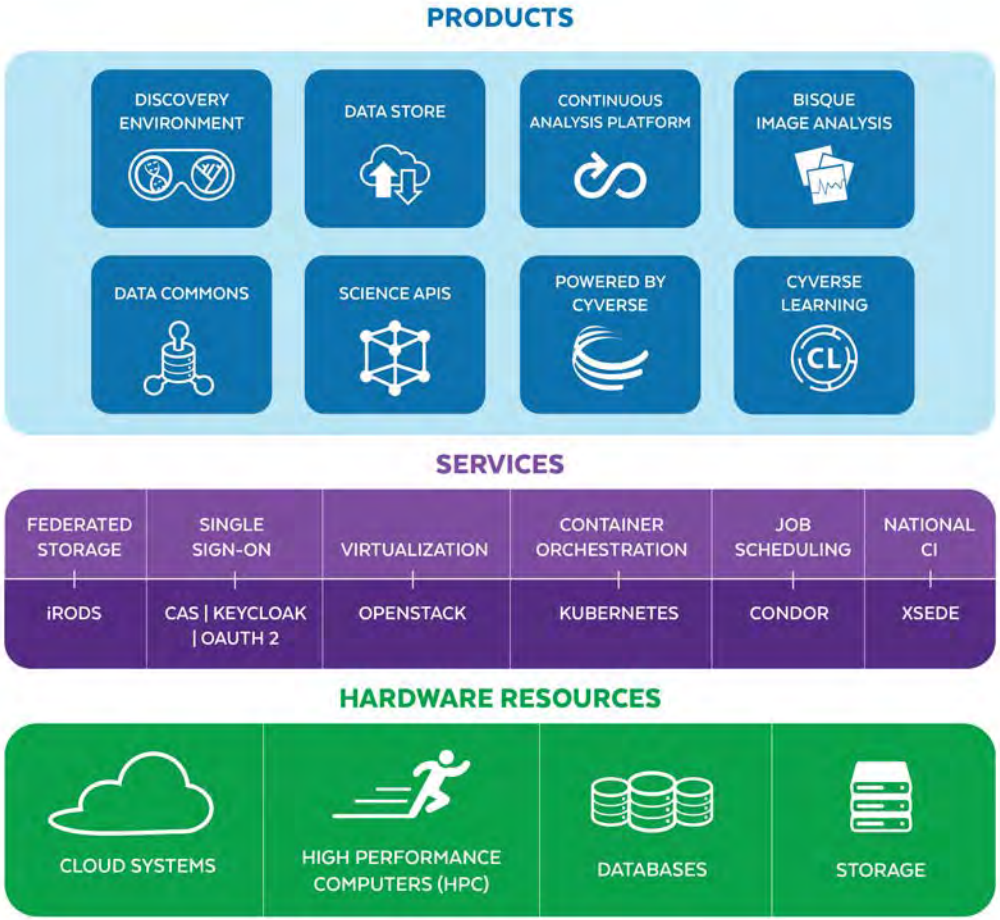
- **Species**: Duckweed (S. polyrhiza ~150 mb)
- **Relevance**: Biofuels and climate change
- **Partners**: CSHL/HudsonAlpha
- **Results:** ~7Gb DNA sequence and partial assembly



Two species of duckweeds (*Wolffia globosa* and *Spirodela polyrhiza*) taken in Waimanalo, Hawai'i by Eric Guinther

2022

- **Species**: Jamaican Broom (C. glandulosa v. mirabilis ~unknown mb)
- **Relevance**: Endemic/endangered
- **Partners**: U. Puerto Rico
- **Results:** +9GB and counting; in-progress

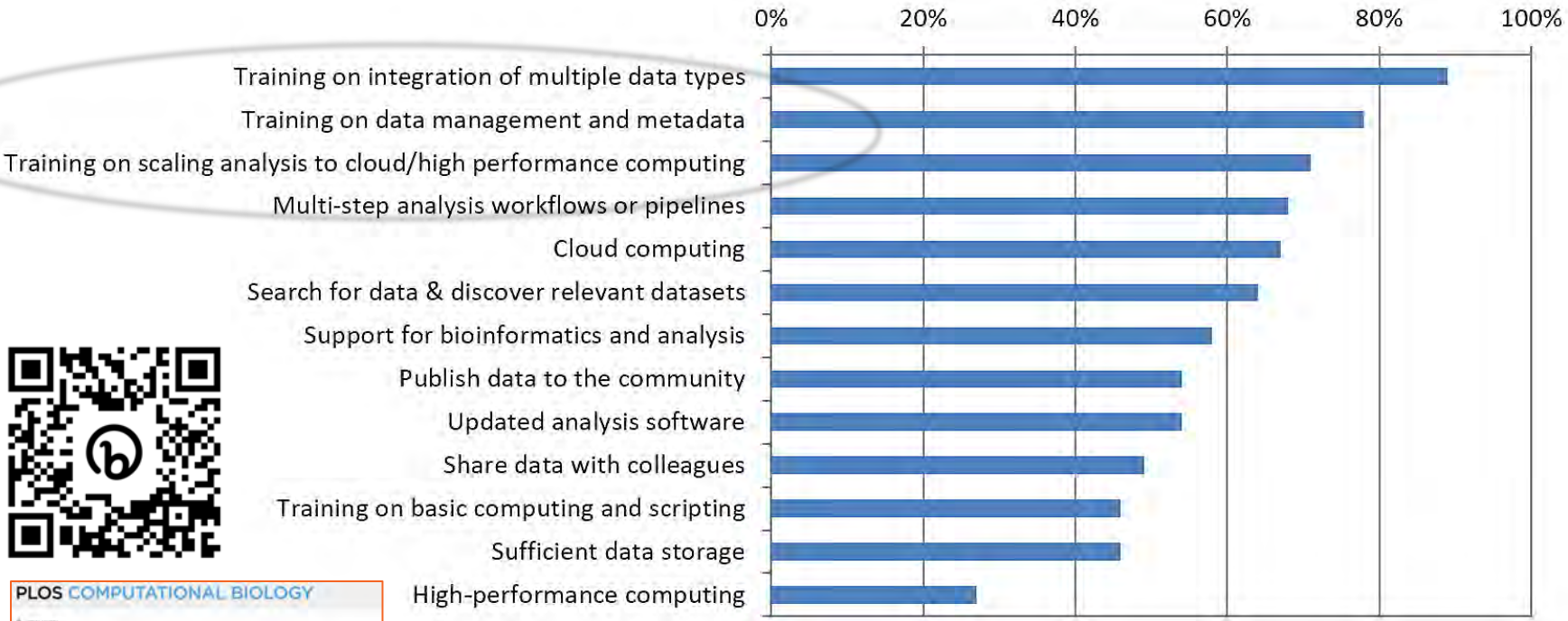


CYVERSE™

- NSF-Subsidized cyberinfrastructure resources
- Easy-to-use interfaces
- Facilitates national/international collaboration (CyVerse UK, CyVerse Austria)

DBI-0735191, DBI-1265383, DBI-1743442

Training is the biggest need



PLOS COMPUTATIONAL BIOLOGY

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction

Jason J. Williams, Jennifer C. Drew, Sebastian Galindo-Gonzalez, Srebrenka Robic, Elizabeth Dinsdale, William R. Morgan, Eric W. Triplett, James M. Burnette III, Samuel S. Donovan, Edison R. Fowlks, Anya L. Goodman, Nealy F. Grandgenett, Carlos C. Goller, [...], Mark A. Pauley [view all]

Published: November 18, 2019 • https://doi.org/10.1371/journal.pone.0224288

NIBLSE
Bioinformatics in Life Sciences Education

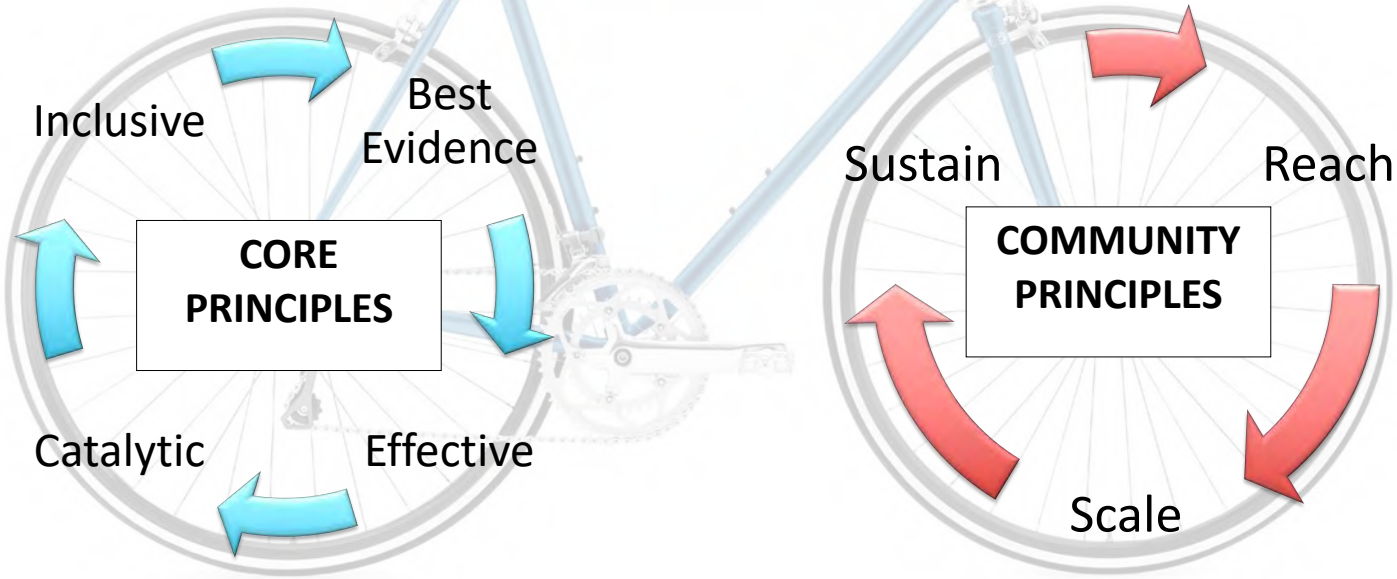
Effective and Inclusive Career-spanning Training

bikeprinciples.org

This material is based upon work supported by the National Science Foundation under DRL/EHR:2027025.

Cold Spring Harbor Laboratory

“The bicycle” – good for going far



A. Professionalize the training of short-format training instructors and instructional designers

Doctor studying a textbook

E. Deploy short-format training to counter inequity

Judge in robe fighting for justice

Building a community of practice in training

LIFESCITRAINERS.ORG

ACCELERATING SCIENTISTS

Why short-format training?

In many areas of the life sciences new technologies and approaches (especially, but not only computational ones) are changing rapidly. It's not possible for formal training (undergraduate/graduate) to keep pace, but short-format training can fill these gaps. Short-format training comes with its own set of challenges, and this community works together to address them.

How I Teach Life Scientists…by Using Reproducible and Scalable Learning Environments

The combination of Docker + cloud computing service enables a teacher to create a highly scalable and flexible learning environment

Continue reading »

How I Teach Life Scientists…to Build Reproducible, Scalable Workflows with Nextflow

The term "reproducible research" has been used to describe the idea that a scientific publication should be distributed along with all the raw data and metadata used in the study, all the code and/or computational notebooks needed to produce results…

Continue reading »

March 2022: Community Discussion – The Return to In-Person Training

Many instructors have or shortly will be running their first in-person events since the pandemic, what will change?

Continue reading »

"The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn"
– A. Toffler

https://bit.ly/grei-jan-2023

Thanks!

Cold Spring Harbor Laboratory
DNA Learning Center

NSF

CYVERSE™

DBI-0735191, DBI-1265383, and DBI-1743442

Bicycle Principles: This material is based upon work supported by the National Science Foundation under DRL/EHR:2027025. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

@JasonWilliamsNY

williams@cshl.edu

BACKUP SLIDES

OUR INSTITUTIONAL MEMBERS

California State University, East Bay, University Libraries
Chan Zuckerberg Initiative
Claremont Graduate University
Claremont McKenna College
Columbia University Libraries
Desert Research Institute
Harvey Mudd College
Indian Institute of Technology Dhanbad
Keck Graduate Institute
Kyoto University
Lawrence Berkeley National Laboratory (LBNL)
Macquarie University Library
Michigan State University Libraries
Montana State University Library
Netherlands Institute of Ecology (NIOO-KNAW), Library and Information Services
New Mexico State University, Research Cores Program
North Carolina State University, University Libraries
Ohio State University, University Libraries
Pitzer College
Pomona College
Scripps College
Stanford University, School of Medicine, Lane Medical Library
Texas Tech University, IT Technology Support
The University of New South Wales Library
University of California, Berkeley (UCB)
Univer
Califor
Univer
(UCD)
Univer
(UCI)
Univer
Angel
University of California, Merced (UCM)
University of California, Riverside (UCR)
University of California, San Diego (UCSD)
University of California, San Francisco (UCSF)
University of California, Santa Barbara (UCSB)
University of California, Santa Cruz (UCSC)
University of Illinois, University Library
University of Minnesota Libraries
University of New Mexico,

OUR PUBLISHER MEMBERS

American Academy for the Advancement of Science
American Genetic Association
American Society of Naturalists
American Society of Plant Taxonomists
Association for Tropical Biology and Conservation
British Ecological Society
Cambridge University Press
eLife
European Society for Evolutionary Biology
Hindawi
John Wiley & Sons, Inc
Nordic Society Oikos
Oxford University Press
The Paleontological Society
The Royal Society
Society for the Study of Evolution
Society of Systematic Biologists
The University of Chicago Press

Researchers, institutions and publishers invest in Dryad for the curation, open sharing and reuse of research data in all fields.

Principles of Open Scholarly Infrastructure

DRYAD

Not-for-profit and online since 2008. Learn more: datadryad.org

Data and metadata are:

- ➔ **Curated by our expert team to ensure suitability for publishing and to facilitate discovery and reuse**
- ➔ **Permanently stored in our CoreTrustSeal repository**
- ➔ **Published under a Creative Commons Public Domain (CC0) license**
- ➔ **Accessible via our open API**

- ➔ Serving all research domains
- ➔ Leader in research data
- ➔ Interconnected

How it works

Image CC0 https://doi.org/10.5061/dryad.xd2547dd5

“Data should be of sufficient quality to validate and replicate research findings”

* https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html#_ftn8

Dryad supports

Data publishing at Dryad:

1. **Ensures metadata quality**
2. **Increases discoverability of data**
3. **Connects data with other research outputs**
4. **Promotes data citation**
5. **Makes data count**

DRYAD

Not-for-profit and online since 2008. Learn more: datadryad.org

Dryad is a founding member of the

DATA CURATION NETWORK
Ethical. Reusable. Better.

Data curation at Dryad:

1. **Ensures data is appropriate and licensed for sharing**
2. **Verifies data is accessible and usable**
3. **Supports authors**