Bringing Open Science in the US into Practice

Christine Kirkpatrick

May 2019

Overview

- **Scope of US Research**
- **Recent policy: progress and results**
  - Agencies, academia, and professional societies
- **Examples of Open Science in action**
  - Distributed, national scale data sharing and infrastructure
  - Adapting EU models to a US audience
- **Conclusion**

Context: US Research Landscape

Producers (data, software, publications)

- Federal agencies (dozens)
- 50 states
- 3,143 counties, municipalities
- $75B (€64B) in research expenditures at US universities
- 1,100 teaching hospitals
- 449 publicly traded biotech companies
  - 600+ private and public biotech companies in San Diego

Source, Wolfram Alpha, NSF Rankings by R&D Expenditures (2017) ,
https://www.statista.com/statistics/197930/number-of-united-states-biotech-companies-by-type/

Sample of Institutions Driving Open Science in the US

Federal Agencies and Offices

- OSTP
- NSF
- NIH
- NIST
- DOE
- NASA
- GSA
- IMLS

Academia and Professional Societies

- NAS
- AGU

Foundations

- Arnold
- Schmidt

the WHITE HOUSE
PRESIDENT BARACK OBAMA

BRIEFING ROOM | ISSUES | THE ADMINISTRATION | 1600 PENN

HOME · BLOG

Holdren Memo (2013)

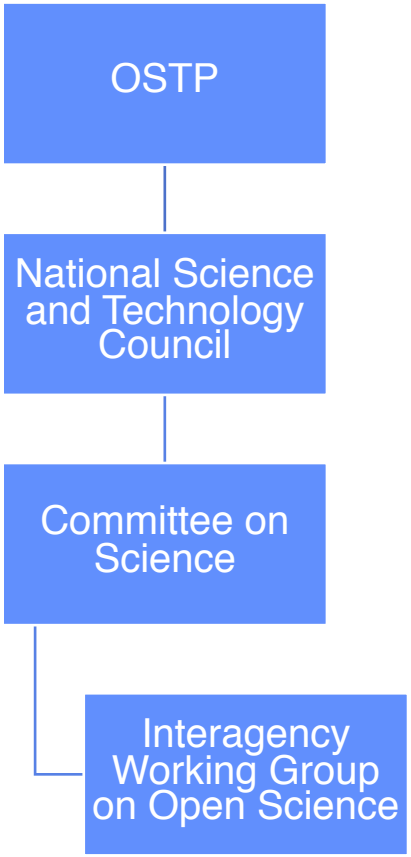
Expanding Public Access to the Results of Federally Funded Research

FEBRUARY 22, 2013 AT 12:04 PM ET BY MICHAEL STEBBINS

Summary: The Obama Administration is committed to the proposition that citizens deserve easy access to the results of research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally funded research freely available to the public—generally within one year of publication.

The Obama Administration is committed to the proposition that citizens deserve easy access to the results of scientific research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research. OSTP has been looking into this issue for some time, soliciting broad public input on

Office of Science Technology & Policy



E. Membership

The following NSTC departments and agencies are represented on the IWGOS:

Department of Agriculture;
Department of Commerce;
Department of Education;
Department of Energy;
Department of Defense;
Department of Health and Human Services (Co-chair);
Department of Homeland Security;
Department of the Interior;
Department of Transportation;
Department of Veterans Affairs;
Environmental Protection Agency;
National Aeronautics and Space Administration;
National Science Foundation (Co-chair);
Office of the Director of National Intelligence;
Smithsonian Institution; and
U.S. Agency for International Development.

Source: White House Archives

US Agencies

National Science Foundation

- **Snapshot**
  - $8.1B (FY2019)
  - 27% basic research at universities
  - Major source for mathematics, computer science, social sciences
- **Open science in practice**
  - NSF PAR
  - Biosketch format
  - Deposit fees allowable
  - Several funded projects (72+ active)

Source: https://www.nsf.gov/about/glance.jsp

NSF Public Access Repository (NSF-PAR)

A partnership with the Department of Energy,
Office of Scientific and Technical Information

National Science Foundation
WHERE DISCOVERIES BEGIN

Explore scholarly publications in the NSF Public Access Repository

"gravitational waves"

Find

+ Advanced Search

Home / Search Results / Page 1 of 92

Search for: "gravitational waves"

Sort by Relevance

« Prev | Next »

Total Results 917

Publicly Available Full Text 845

Citation Only 72

Filtered Results

Filter Results

Filter by Author

Save Results

Excel

CSV

XML

Have feedback or suggestions for a way to improve these results?
Let us know !

Note: *When clicking on a Digital Object Identifier (DOI) number, you will be taken to an external site maintained by the publisher. Some full text articles may not yet be available without a charge during the embargo (administrative interval).*

Some links on this page may take you to non-federal websites. Their policies may differ from this site.

1. *Optical scattering measurements and implications on thermal noise in **Gravitational** Wave detectors test-mass coatings*
   doi: https://doi.org/10.1016/j.physleta.2017.05.050
   Glover, Lamar ; Goff, Michael ; Patel, Jignesh ; Pinto, Innocenzo ; Principe, Maria ; Sadecki, Travis ; Savage, Richard ; Villarama, Ethan ; Arriaga, Eddy ; Barragan, Erik ; et al (August 2018, Physics letters. A)

   > *Photographs of the LIGO **Gravitational** Wave detector mirrors illuminated by the standing beam were analyzed with an astronomical software tool designed to identify stars within images, which extracted hundreds of thousands of point-like scatterers uniformly distributed across the mirror surface, likely distributed through the depth of the coating layers. The sheer number of the observed scatterers implies a fundamental, thermodynamic origin during deposition or processing. These scatterers are a possible source of the mirror dissipation and thermal noise foreseen by V. Braginsky and Y. Levin, which limits the sensitivity of observatories to **Gravitational Waves**. This study may point the way* more »

   Free, publicly-accessible full text available August 25, 2019

2. *A Study of **Gravitational** Wave Memory and Its Detectability With LIGO Using Bayesian Inference*
   Doane, Jillian ; Weinstein, Alan ; Kanner, Jonah (July 2018, LIGO Laboratory Summer 2018 Undergraduate Research)

   > *The detectable component of **gravitational waves**, known as the oscillatory waveform, is predicted to have a smaller, lower frequency counterpart called the memory: a permanent warping of space-time. The memory component is low-frequency (below the*

National Institutes of Health

Snapshot

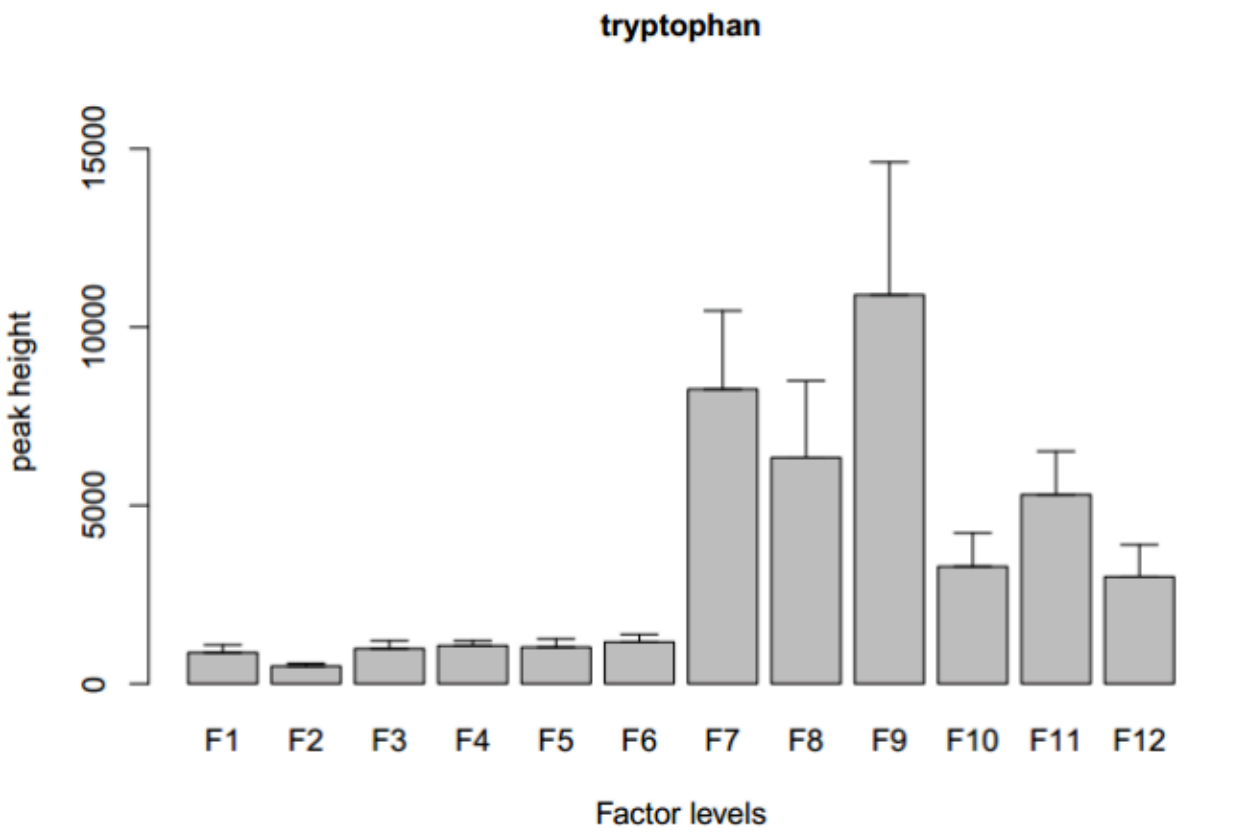
- **$39.2B (€28.5B) awarded in research grants per year**
- **10% budget goes to NIH scientists (6,000)**
- **"…fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability."**

Open science in practice

- **FAIR plan**
- **Data commons / Data repositories**

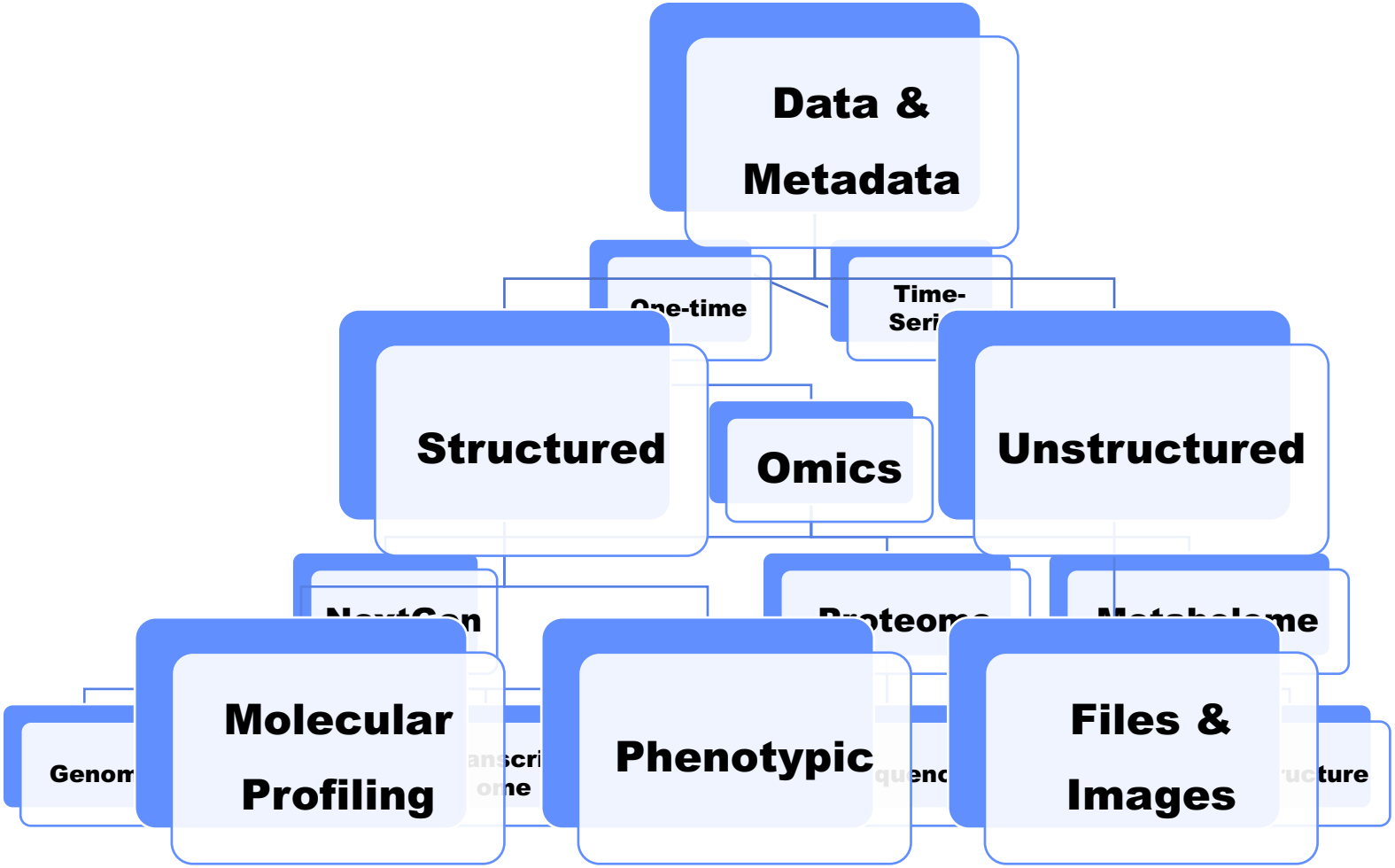
Source: About NIH

Bar graph of mean values for each factor level



Factor level F1:	Compartment:cytosol | Minutes:0 | Skeletal Muscle Treatment:palmitic acid 2.4 uM
Factor level F2:	Compartment:cytosol | Minutes:0 | Skeletal Muscle Treatment:palmitic acid 9.6 uM
Factor level F3:	Compartment:cytosol | Minutes:0 | Skeletal Muscle Treatment:palmitoyl carnitine 9.6 uM

Source: Metabolomics Workbench Tutorial, metabolomicsworkbench.org funded by NIH #1U2CDK119886-01 .



Source: Dr Shankar Subramaniam

National Institute of Standards and Technology

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Snapshot

- $985M (FY 2019)
- Core competencies: measurement science, rigorous traceability, development and use of standards

Open science in practice

- Thought leadership
- Open formats for equipment
- Workshop and community support

Department of Energy

Snapshot

- $30.6B annual budget
- $5.4B R&D
- Nuclear (safety), energy

EXASCALE COMPUTING PROJECT

Open science in practice

- Community databases, e.g. ARM, materials
- Software interoperability

National Aeronautics and Space Administration (NASA)

Snapshot

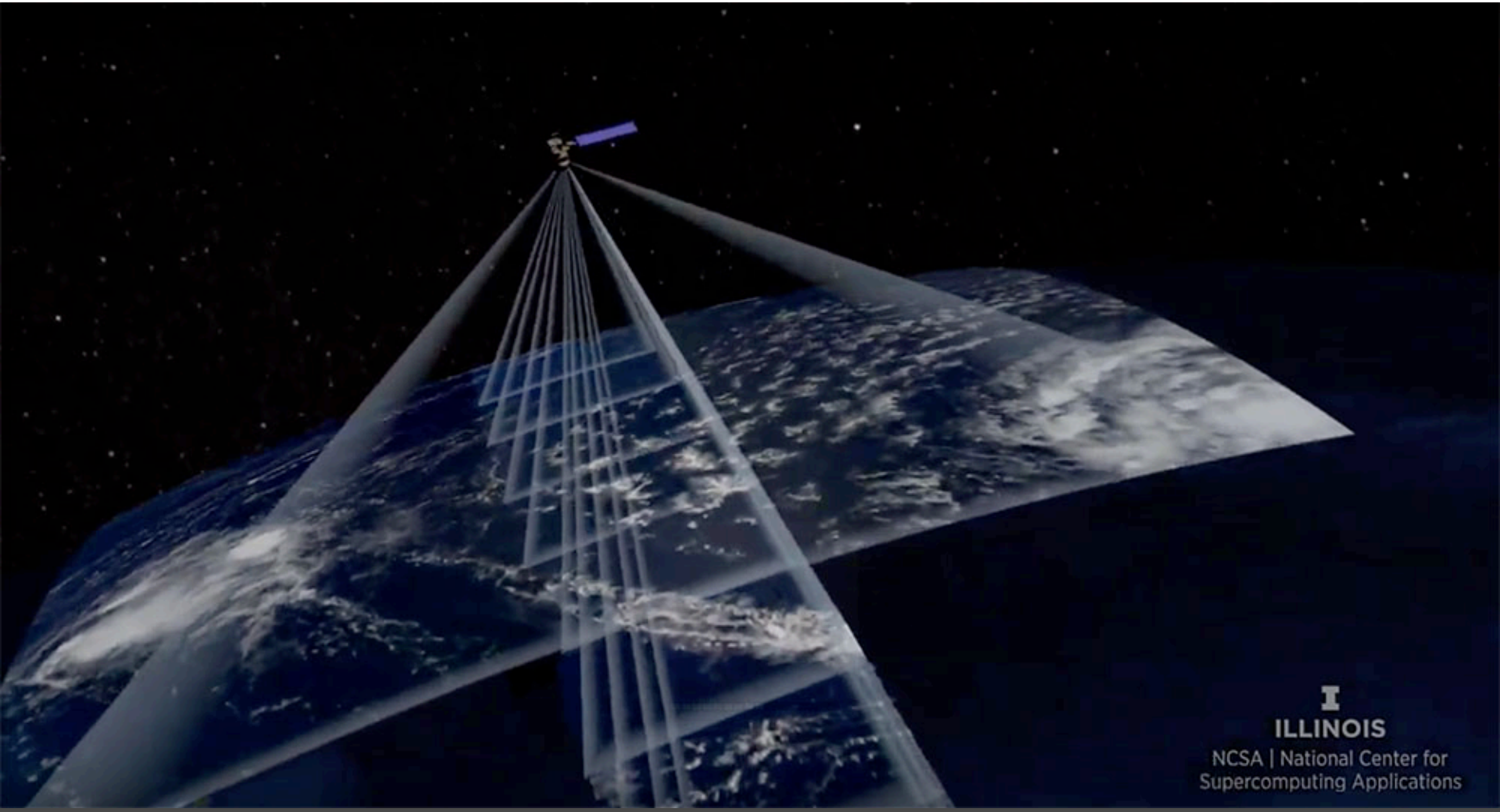
- **$21.5B**
- **"Open-access culture"**

Open science in practice

- **% budget for data (management)**
- **Long history of open data, data (active) archive centers (DAAC)**
- **Innovative researchers/research projects**

Source: NASA Budget Brief, NASA Plan for Increasing Access to the Results of Scientific Research, https://earthdata.nasa.gov/earth-science-data-systems-program/competitive-programs/access/terra-data-fusion-products

Terra Fusion Project (Larry DiGiralomo)



Source: https://earthdata.nasa.gov/earth-science-data-systems-program/competitive-programs/access/terra-data-fusion-products

National Academy of Sciences, Engineering, and Medicine

1. Committee Toward an Open Science Enterprise

- Open Science by Design: Realizing a Vision for 21st Century Research

2. Roundtable on Aligning Incentives for Open Science

National Academies of Sciences, Engineering, and Medicine 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116.

Other Notable Contributions to Open Science in the US

- **General Services Administration**
  - Data.gov, code.gov
- **Institute of Museum and Library Services (IMLS)**
- **American Geophysical Union**
- **Arnold Foundation**
  - Open Science Framework
- **Schmidt Foundation**
  - Open Storage Network

code.gov

code_ ABOUT BROWSE PROJECTS FEDERAL AGENCIES DEVELOPERS SOCIAL

There are 70 open tasks

Filter

Federal Agency

- [ ] Consumer Financial Protection Bureau
- [ ] Department of Defense
- [ ] Department of Energy
- [ ] Department of Health and Human Services

Show more

Skill Level

- [ ] Beginner
- [ ] Intermediate
- [ ] Advanced

Explore Open Tasks

multilingual support?

Agency: General Services Administration **Last Updated:** 12/25/2018

Languages: Not Available **Type:** Enhancement **Skill Level:** Intermediate **Effort:** Medium

Integrate Accessibility Testing into CircleCI Pipeline

Agency: General Services Administration **Last Updated:** 4/30/2019

Languages: Not Available **Type:** Enhancement **Skill Level:** Intermediate **Effort:** Medium

Wagtail: image and text 25/75 requires alt image tag twice

Agency: Consumer Financial Protection Bureau **Last Updated:** 4/18/2018

Languages: Not Available **Type:** Not Available **Skill Level:** Advanced **Effort:** Medium

OSF HOME ▾ Search Support Donate Sign Up

PresQT Data and Software Preservation... Files Wiki Analytics Registrations

Public

PresQT Data and Software Preservation Quality Tool Project

Contributors: John Wang, Sandra Gesing, Rick Johnson, Natalie Meyers, Jeffrey R. Spies, David Minor, Markus Krusche

Affiliated institutions: Center For Open Science, University of Notre Dame

Date created: 2016-05-30 05:09 PM | Last Updated: 2018-12-20 07:49 AM

Identifiers: DOI 10.17605/OSF.IO/D3JX7 | ARK c7605/osf.io/d3jx7

Category: Project

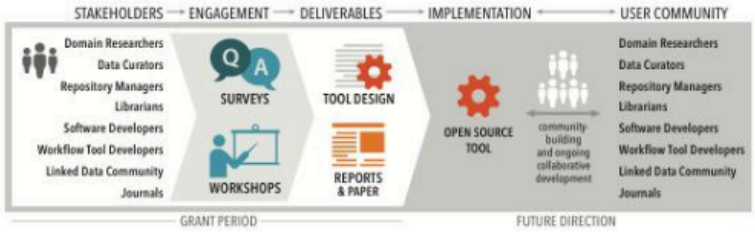
Description: The goal is to collaboratively design interoperable and repository agnostic data and software preservation quality tools.

License: CC-By Attribution 4.0 International

Wiki

Research Data & Software Preservation Quality Tool Effort

The Goal: is to collaboratively design an interoperable and repository agnostic Data and Software Preservation Quality Tool.



Citation

Components

Partner Meeting January 28-29, 2019

Anderson, Branco, Brower & 24 more

Documents for and from the Protoyping Partner Meeting January

Implementation Effort

Brower, Gesing, Johnson & 20 more

Start managing your projects on the OSF today.

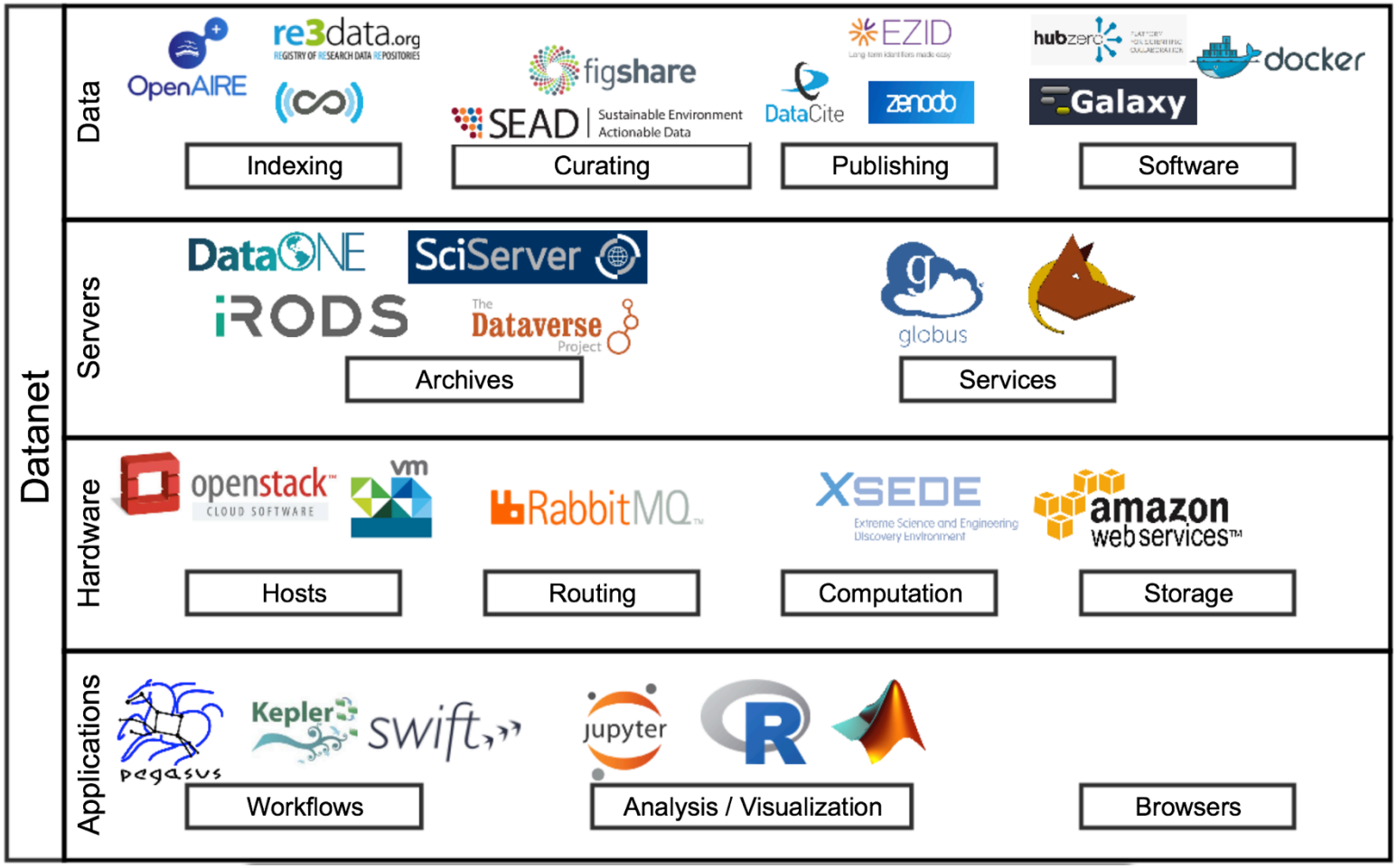
SDSC SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY OF CALIFORNIA

Made possible in part by the Institute of Museum and Library Services, IMLS, awards LG-72-16-0122-16 and LG-70-18-0082-18.

US National Data Service

Where We Are



Source:: Kenton McHenry, NCSA

Research is More than Publications

- **Equal Partners: Pubs, Data, Analysis**
- **Transparency and access to methods**
- **FAIRR**
  - Reusable
  - *Reproducible*

U.S. National Data Service

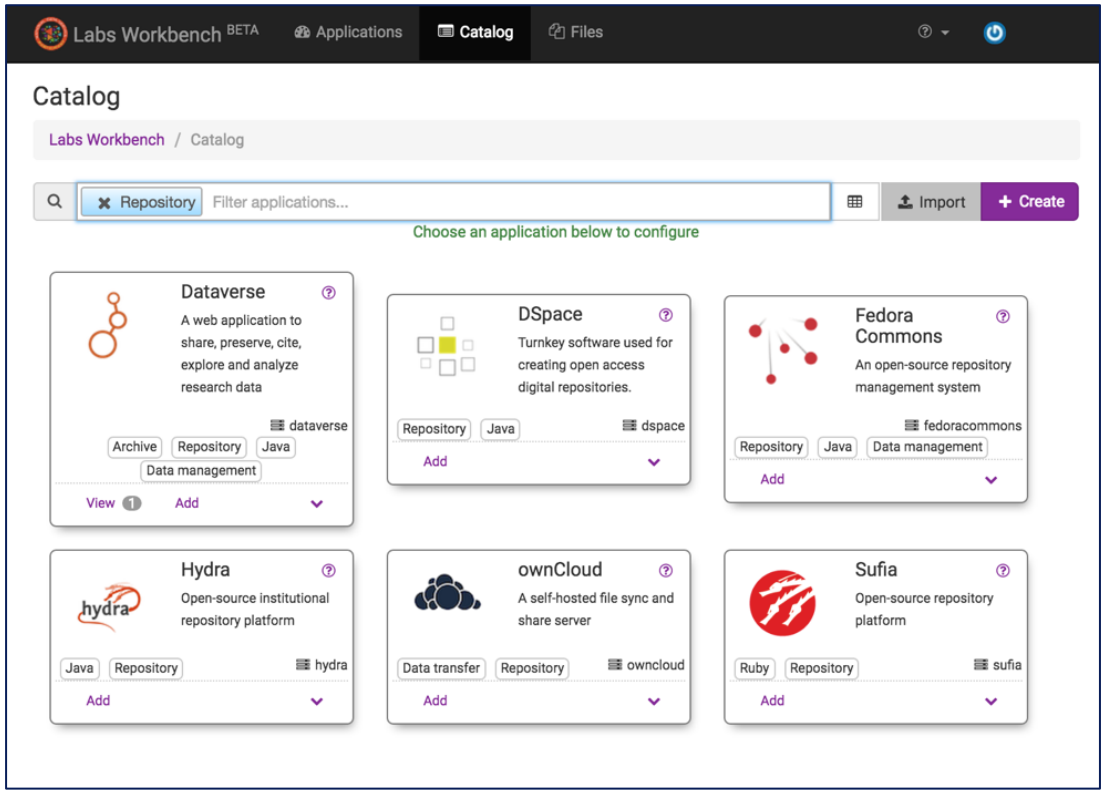
National effort to bring together infrastructure supporting the *publication*, *discovery*, and *reuse* of data

→ From the Internet to the "Datanet"

1. **Large-scale Data Service Interoperability**
   - Distributed cloud and compute
   - Innovation in the gaps:
     services, software, integration
2. **Incubator of Data Projects & Pilots**
   - Quick start sandbox
   - Choose services based on features
     (not time to install)
3. **Training Platform**

NDS Labs Workbench: Tools-centric

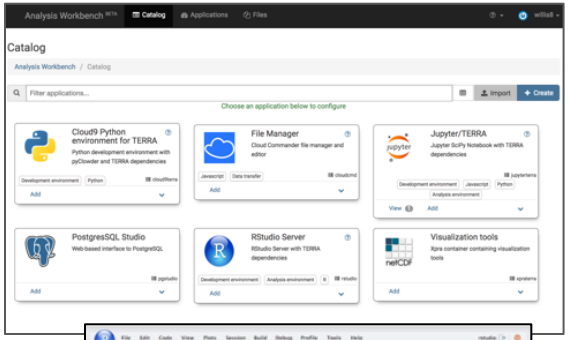
- Open source project. Initiative since January 2016.
- Public beta

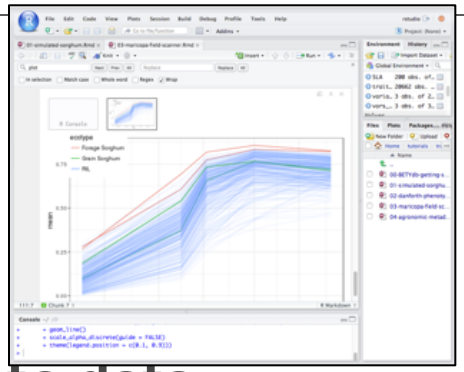


https://www.workbench.nationaldataservice.org

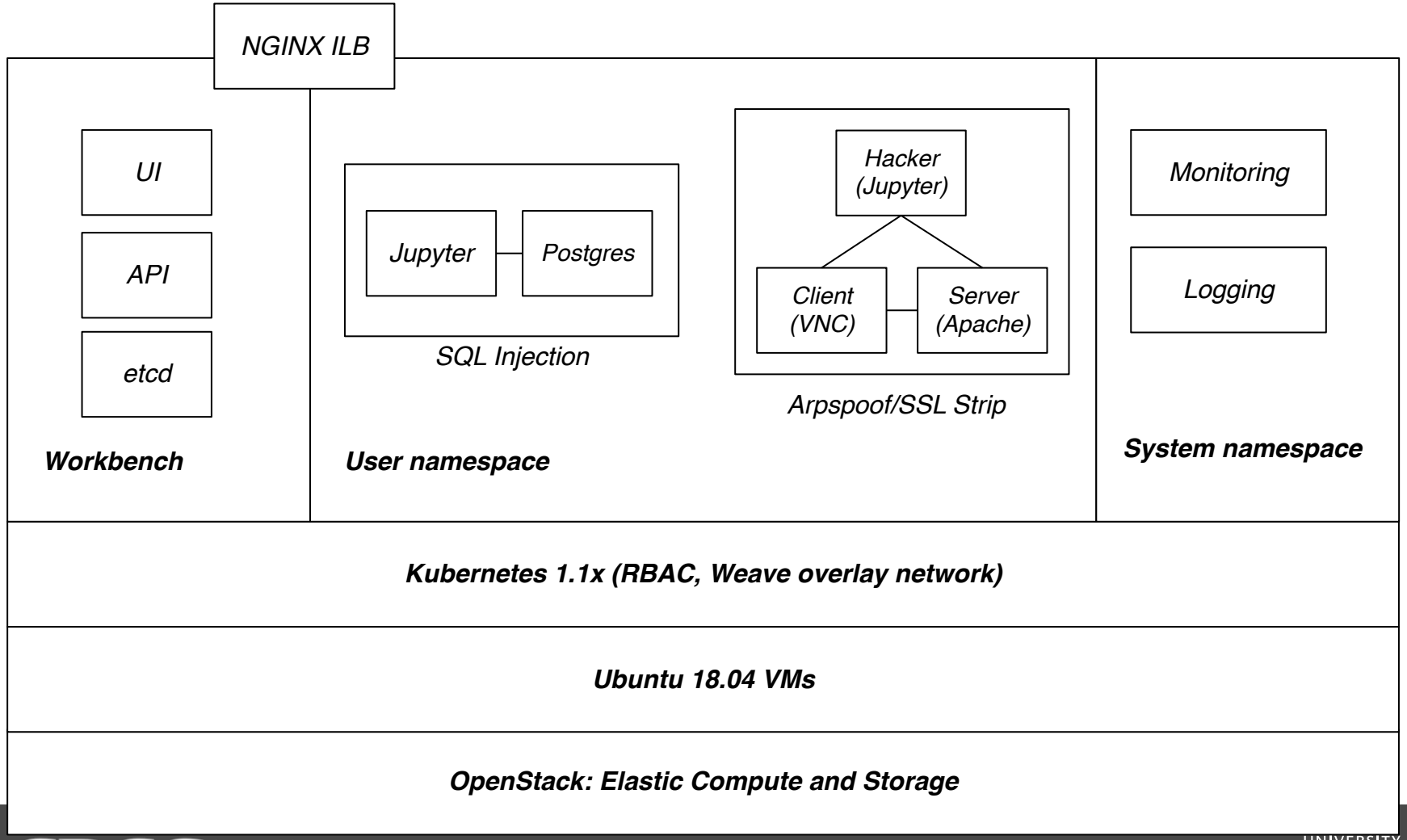
Use case: TERRA-REF

- **High-throughput indoor and outdoor sensor platforms, UAV and field data, large-scale genome sequencing**
- **Petabyte scale data storage and computing pipeline**
  - Data processing workflows
  - Raw and derived data
  - Data sharing and re-use
- **Nationwide, multi-institution collaboration**
  - Researchers, data scientists, and software developers
- **Labs Workbench for remote, interactive access to data**





CHEESE Technologies/Architecture



SDSC SAN DIEGO SUPERCOMPUTER CENTER
UNIVERSITY OF CALIFORNIA
Source: Baijian "Justin" Yang & Rajesh Kalyanam, Purdue
This material is based upon work supported by the National Science Foundation under NSF grant # 1820573.

DataDNS: Data-centric Portal

DataDNS | Datasets | Register Dataset

Registered Data Sets

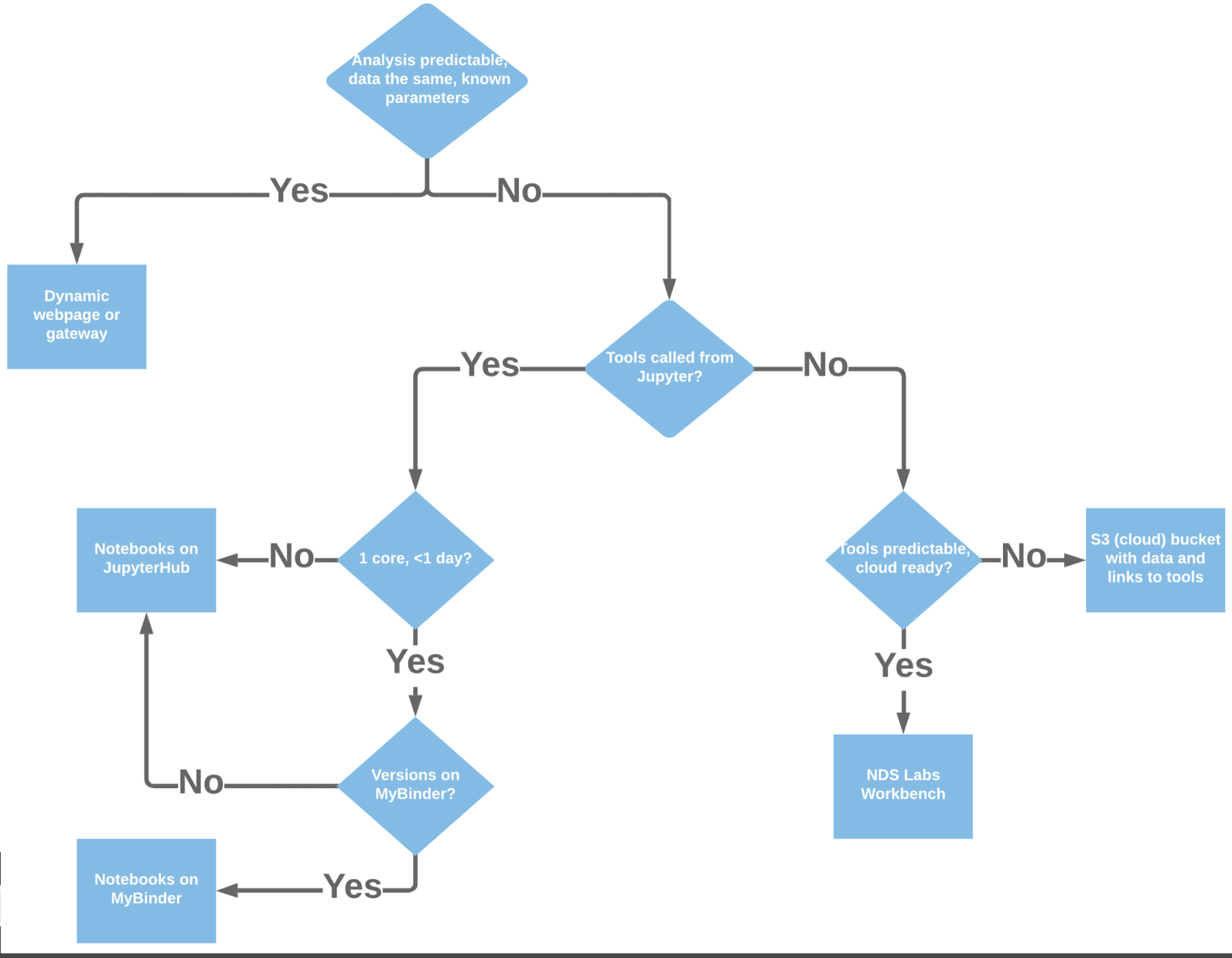
Search...

Dataset	Publications	Location	Launch Notebook	Show Metrics
Renaissance Simulations O'Shea, Brian (oshea@msu.edu); Wise, John; Xu, Hao; Norman, Michael Cite Dataset More Information	• O'Shea, B. W., Wise, J. H., Xu, H., & Norman, M. L. (2015). PROBING THE ULTRAVIOLET LUMINOSITY FUNCTION OF THE EARLIEST GALAXIES WITH THE RENAISSANCE SIMULATIONS. The Astrophysical Journal, 807(1), L12. doi:10.1088/2041-8205/807/1/l12 • Ahn, K., Xu, H., Norman, M. L., Alvarez, M. A., & Wise, J. H. (2015). SPATIALLY EXTENDED 21 cm SIGNAL FROM STRONGLY CLUSTERED UV AND X-RAY SOURCES IN THE EARLY UNIVERSE. The Astrophysical Journal, 802(1), 8. doi:10.1088/0004-637x/802/1/8 More	SDSC SAN DIEGO SUPERCOMPUTER CENTER		
Dark Sky Simulations Warren, Michael; Friedland, Alexander; Holz, Daniel; Skillman, Samuel; Sutter, Paul; Turk, Matthew (mjturk@illinois.edu); Wechsler, Risa Cite Dataset More Information	• S. W. Skillman, M. S. Warren, M. J. Turk, R. H. Wechsler, D. E. Holz, P. M. Sutter. Dark Sky Simulations: Early Data Release. • Warren, M. S., Friedland, A., Holz, D. E., Skillman, S. W., Sutter, P. M., Turk, M. J., & Wechsler, R. H. (2014). Dark Sky Simulations Collaboration. ZENODO. https://doi.org/10.5281/zenodo.10777	yt		
Magnetohydrodynamic Turbulence Simulations Mösta, Philipp (pmoesta@berkeley.edu) Cite Dataset More Information	• Mösta, P., Ott, C. D., Radice, D., Roberts, L. F., Schnetter, E., & Haas, R. (2015). A large-scale dynamo and magnetoturbulence in rapidly rotating core-collapse supernovae. Nature, 528(7582), 376–379. doi:10.1038/nature15755	NCSA		

JupyterHub and MyBinder

- **JupyterHub**
  - A multi-user Hub, spawns, manages, and proxies multiple instances of the single-user Jupyter notebook server.
  - Requires your own infrastructure.
- **MyBinder**
  - A Git repository that contains:
    - Code that other people can run (a Jupyter Notebook or an R script)
    - The configuration to run the code in a Docker container.
  - Free limited compute resources from www.mybinder.org.

(Data) Platforms Decision Tree



Open Storage Network

US Research Cyber-Infrastructure Today

Computation	Networking	Storage
Shared Resource (XSEDE, PRAC)	*Over 200 universities with 40/100Gb Connectivity*	*Largely Balkanized*
Standardized	*Standardized*	*No Standards Requirement*
NSF-Funded	*NSF-Funded*	*No CI Funding*

Six Prototype Deployment Sites

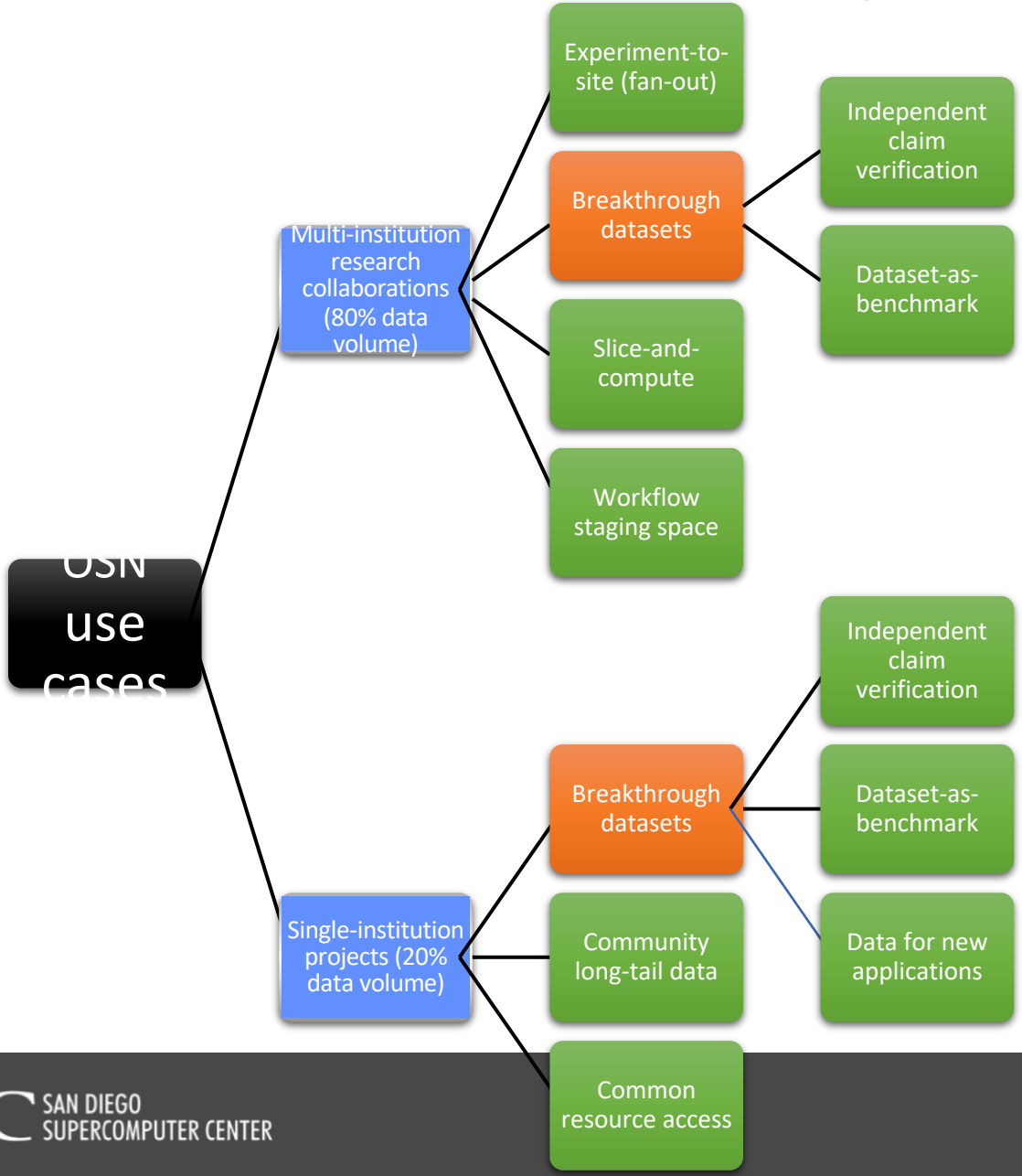
Funded by NSF

Funded by Schmidt Foundation

- Johns Hopkins University
- Massachusetts Green HPC Center
- Northwestern University (Starlight)
- University of CA San Diego (SDSC)
- University of Illinois (NCSA)
- University NC Chapel Hill (RENCI)
- (Atacama Desert, Chile?)

Supported by the Schmidt Foundation and NSF grants #1747552, 1747493, 1747507, 1747490.1747483

OSN Use Case Types



(GO) FAIR US

It Takes a Village to be FAIR

Faulty Assumptions

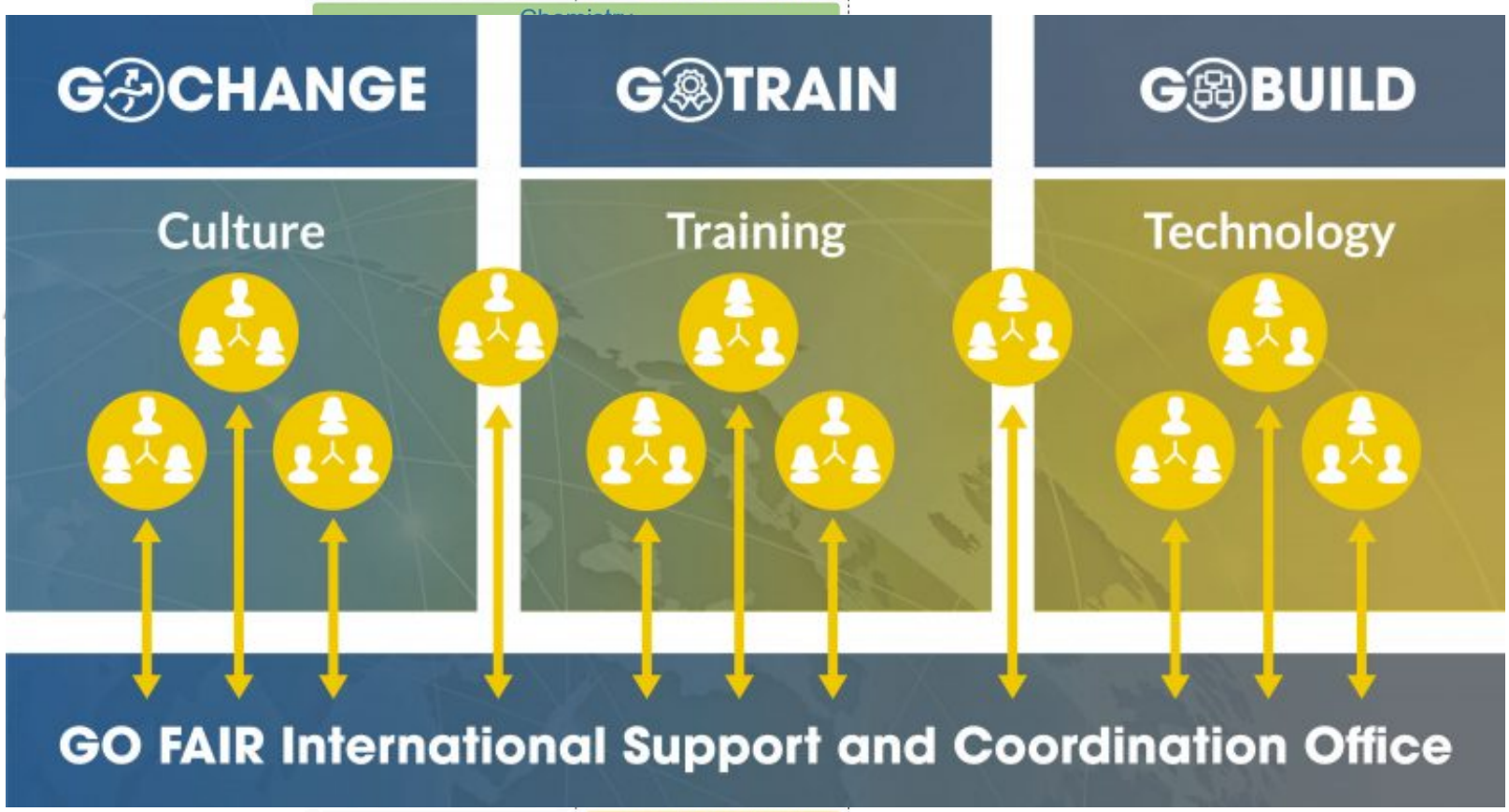
- Curationists will make data FAIR (on their own).
- Security people do all the security work.
- Webmasters make all material accessible.

Partners in FAIR data stewardship:

- Research computing
- Libraries
- Research labs (researchers, postdocs)
- Administrators

Imp rks

GO CHANGE	GO BUILD	GO TRAIN
	Annotation	
	ASTRON	
	BiodiFAIRse	
	Biodiversities	
	C2CAMP	
	CBS (Economics)	



	Sea Data Cloud	
		Season Schools
	(GAIA) System Terre	
		Training Curriculum
		Training Frameworks
	Vaccine IS	

IN Matrix: Converging on metadata/data formats, terminologies

				(DwC-A)/GIF/GeoTIFF/NetCDF/Shapefile/GML/JPEG/MPEG/XLS/XLSX/ODS/ogg/PDF/SVG/TIFF/WAV/XML/TEXT/CSV/Tabular/RTF/RDF/R/SQL/mp3/mp4/avi/			epub, PDF, mobi, and others For data: various (text, tabulated, images, video, audio,...)			FAIRsharing.org		isa-tab, ShapeFiles, OGC formats, IMG...	and more
	Technology	Data Access Protocols (MR/A)	HTTP	HTTP (rfc2616) FTP (rfc959)			HTTP, Z39.50			HTTP, REST		SPARQL, APIs, OGC protocols...	HTTP, FTP (for dataproducts)

GO FAIR US Office

- Train FAIR Data Stewards
  - Train the trainers
- Partnership with Phortos Consultants
  - Training and consulting for local industry
  - Assist with FAIR Data Stewardship Plans
  - Assist organizations/companies to GO FAIR
- Create and harden FAIR tooling
- Extend Implementation Networks (IN) into US

Next GO FAIR Training Opportunity

Who: Those in research labs (researchers, postdocs), research computing, libraries

What: FAIR Data Stewardship Training

When, Where: May 28-31at SDSC, UC San Diego

Cost: $2,500/person
http://tinyurl.com/GOFAIRMay2019

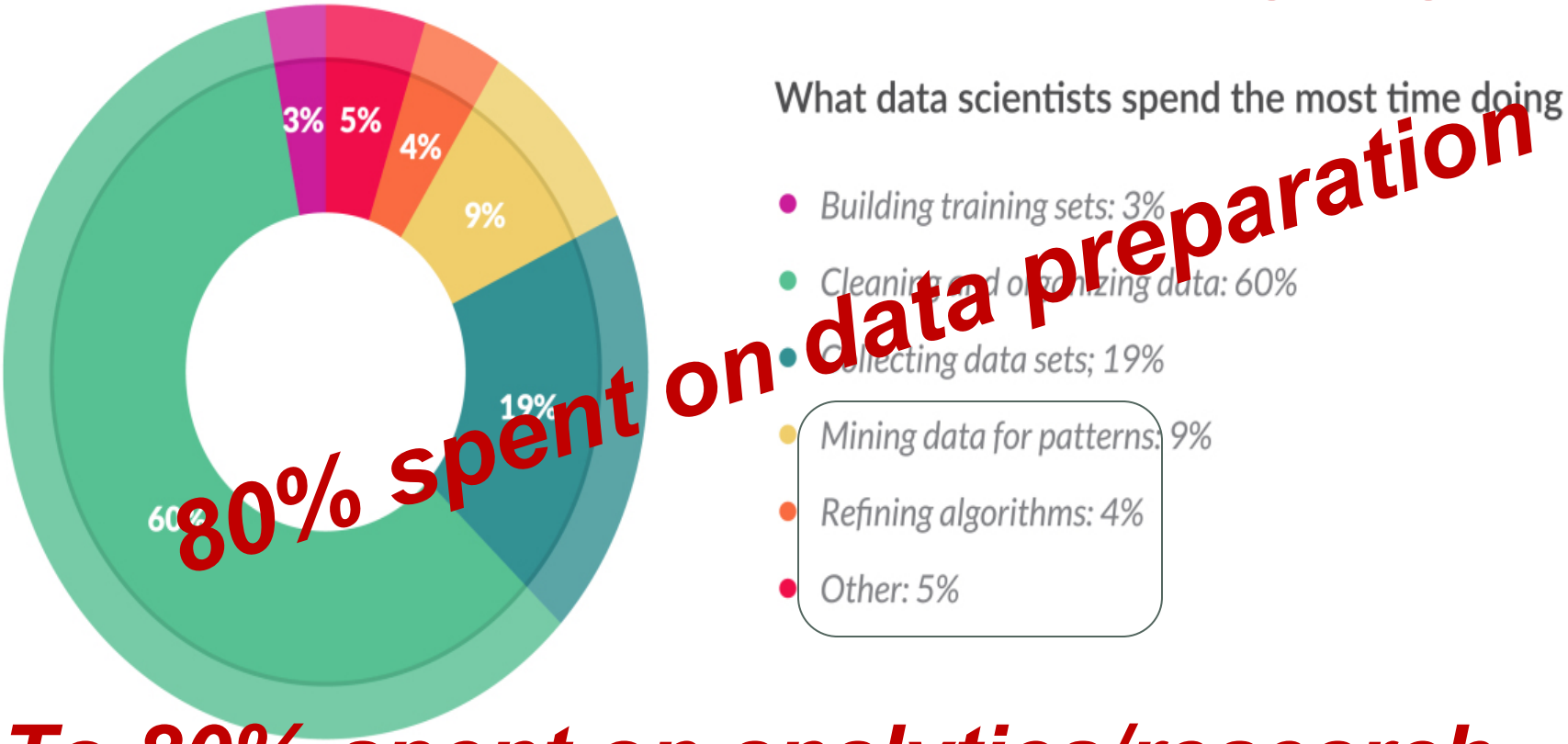
Chopportunities (US)

Challenge + Opportunity

- **Barriers to open science (NAS 2018 report):**
  - Infrastructure costs
  - Subscription-based scholarly communications
  - Lack of incentives
  - Privacy, proprietary barriers
  - Discipline diversity
- **Institutional responsibility to invest in data**
- **[Domain-specific] Scientific advances required**
- **FAIR in industry**

Reframing FAIR in Savings

From 80% spent on data wrangling



To 80% spent on analytics/research

Conclusion

US making good progress

- Open science champions
- Government support
- FAIR imperative → reframe

Path to shaved goat:

- Open science across the research lifecycle
- Scientific advances
- Tooling and training

“…to make the fruits of research and scholarship better and available to all who need or want them.” Berman et al.

christine@sdsc.edu
@SuperChristineK

Christine Kirkpatrick
Division Director, Research Data Services, SDSC
Executive Director, National Data Service