

Structural Biology ~~Data Grid~~ Dataverse

Piotr Sliz,

data.sbgrid.org

Harvard Medical School, Dept. of BCMP
Boston Children's Hospital, Dept. of Pediatrics
SBGrid Consortium



July 11th, 2016

Dataverse Community Meeting 2016
Dataverse Repositories Around the World
Sonia Barbosa

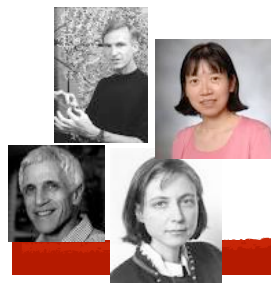
2000-2016: Growth of the community

~300 structural biology groups

(DFCI, BCH, HMS, HU, Tufts, Genzyme)

22 countries

sbgrid.org



2000

2015

The
Dataverse
Project 

CALTECH
Colorado State
Columbia U.
Cornell
Duke University
EMBL Grenoble
ETH Zurich
Genzyme
Harvard
La Trobe U
MIT
NIH
Sloan-Kettering
Novartis
Okinawa Institute of ST
Rice U.
Rockefeller
St. Louis University
Stanford
U of California
U of Iowa
U of MASS.
U of Michigan
U of Toronto
U of Utah
U of Wisconsin-Madison
UT Southwestern
Washington U.
Vanderbilt
Yale

Additional Services Provided by SBGrid in 2016

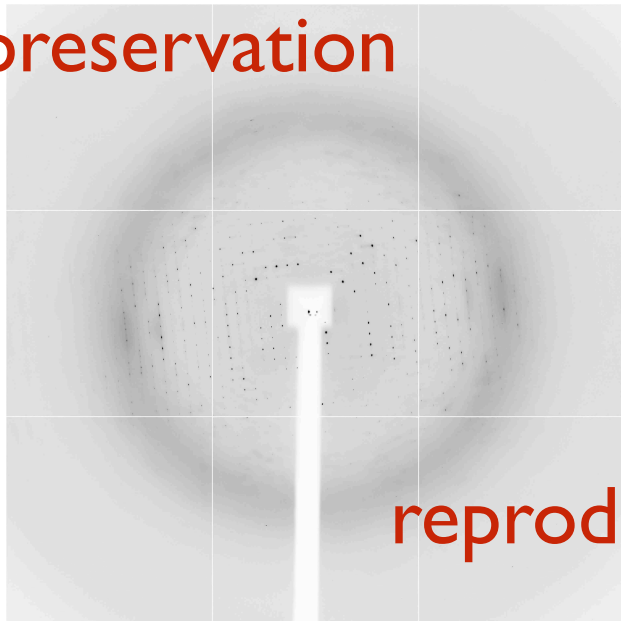
- **Software Support**
 - ▶ (~300 OS, academic, commercial applications)
- **Local Research Computing Infrastructure**
 - ▶ Structural Biology (clusters, storage, workstations)
 - ▶ Lattice Light Sheet Microscopy
- **Access to HPC Resources**
 - ▶ Open Science Grid (Portal)
 - ▶ XSEDE
- **Teaching and Training**
 - ▶ YouTube Program
 - ▶ Structural Biology Training Curriculum (workshops)

Morin, A. et al. Collaboration gets the most out of software.
Elife 2, e01456 (2013).

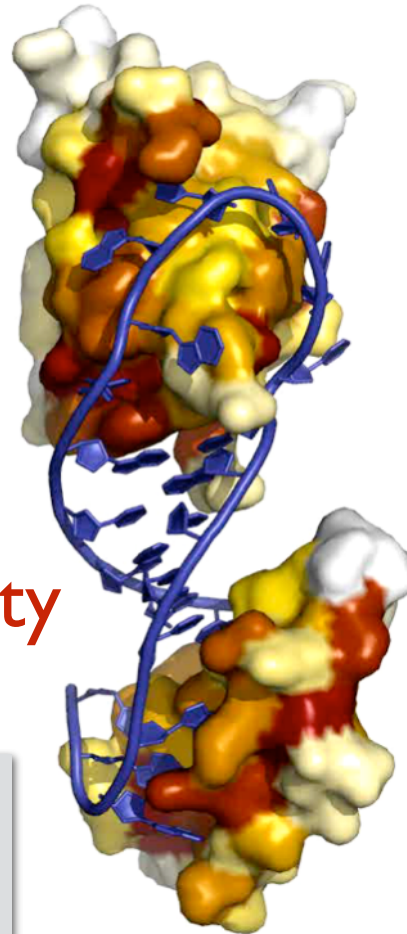
Motivation

Example I: Lin28 crystal structure provides a static interpretation of experimental data and demonstrates limitation of traditional research strategy.

preservation



reproducibility



Nam, et al., 2011 | Cell



MD
Models

- Different software packages (e.g. XDS vs HKL2000: 3D vs 2D profile fitting)
- Different assumptions (e.g. symmetry, mosaicity, radiation damage, number of frames)
- New software packages (e.g. DIALS)
- Improved criteria (e.g. resolution limits such CC1/2, Karplus and Diederichs, 2012, or anisotropic correction)
- New corrections (e.g. data anisotropy)
- Additional features: (e.g. anisotropic diffuse scattering signals)

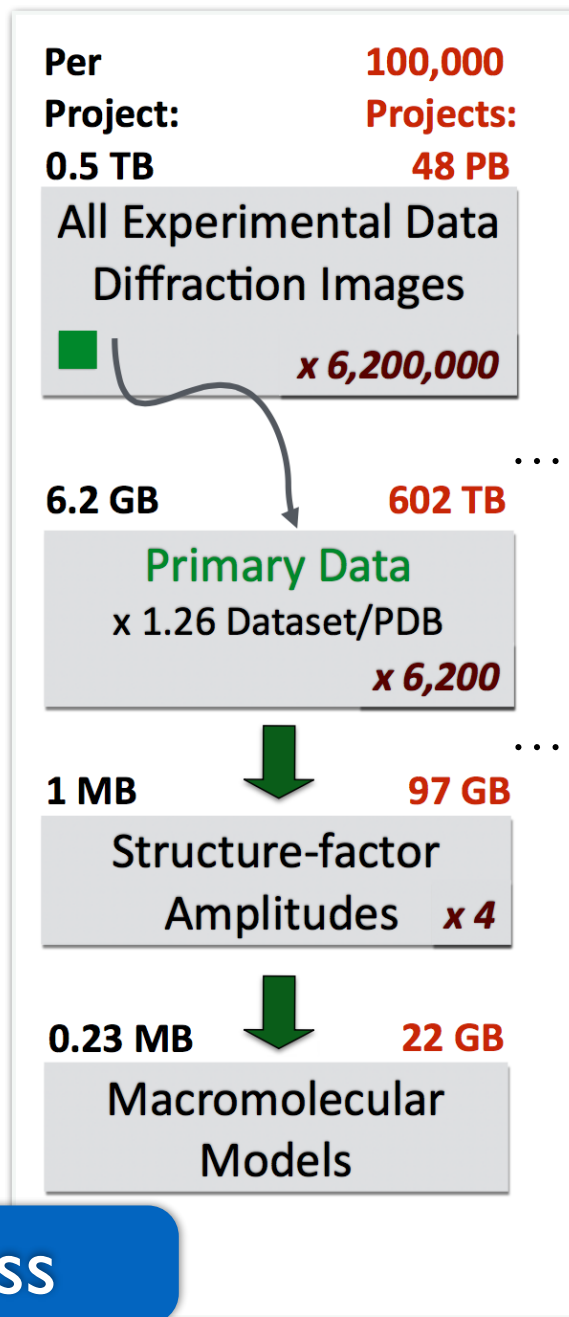


Storage Requirements for Primary Datasets

(based on 100,000 PDB files and average dataset size in SBDG)

SBDG:
110 datasets
~0.5TB

PDB:
100,000 models
0.3 TB



Some of “All Experimental Data” preserved at national synchrotrons e.g. Tardis or Diamond



Primary Diffraction Datasets proposed to be stored on SB Data Grid



Molecular models and reduced datasets are stored in Protein Data Bank

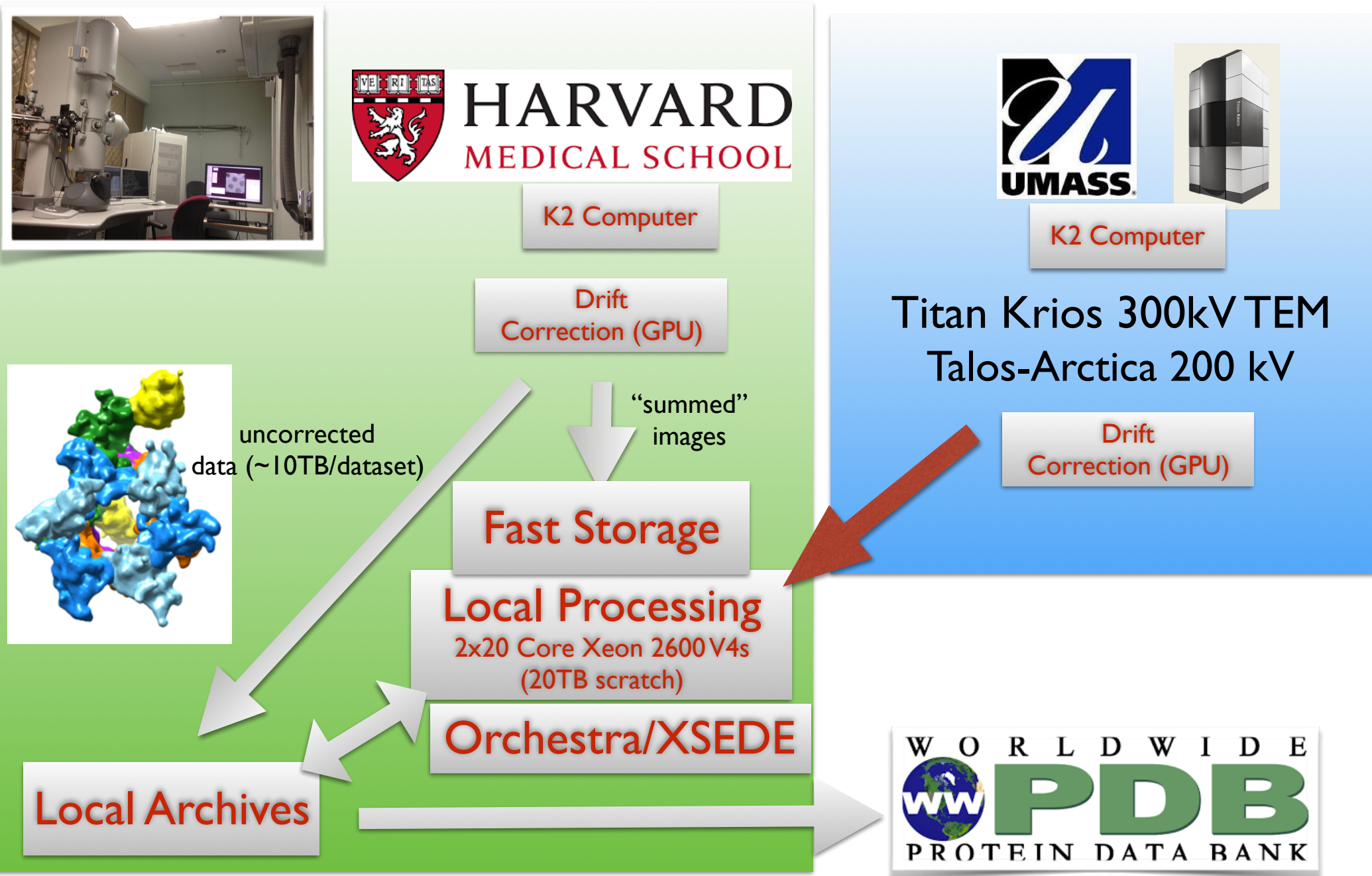


Data Access

Meyer et al, Nature Communications, 2016

Motivation

Example II: EM structure - management of datasets from EM facilities



data.sbgrid.org website

229 Datasets 59 Institutions 196 Structures

Deposit
Share your data with the community. Every dataset deposited with SBDB receives a unique DOI and its own landing page here in the Data Grid.

Explore
Browse all published datasets and download via rsync. Only SBGrid affiliates can download but everyone can view.

Cite
Give credit to the data used in your research. Every dataset published with SBDB generates its own citation to be used within manuscripts.

Lab Collections Institutional Collections

SBGrid DATA BANK

For Depositors Data ▾ About ▾ Get Help ▾

We support publication of X-ray diffraction datasets. All visitors can access the following Laboratory and Institutional Collections. All SBGrid affiliates are invited to deposit datasets.

Lab/Institutional Collections All Datasets

Datasets: 117 Lab/Institutional Collections: 50 Next Update: Friday 5pm

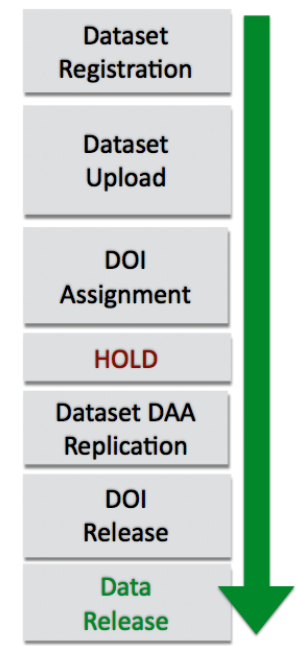
- Anderson Laboratory
Yale University School of Medicine
- Baxter Laboratory
Yale University
- Boggon Laboratory
Yale University School of Medicine
- Bonvin Laboratory
Utrecht University
- Brett Laboratory
Washington U. School of Medicine
- Buschiazio Laboratory
Institut Pasteur de Montevideo

SBGrid DATA BANK

For Depositors ▾ Data ▾ About ▾ Get Help ▾

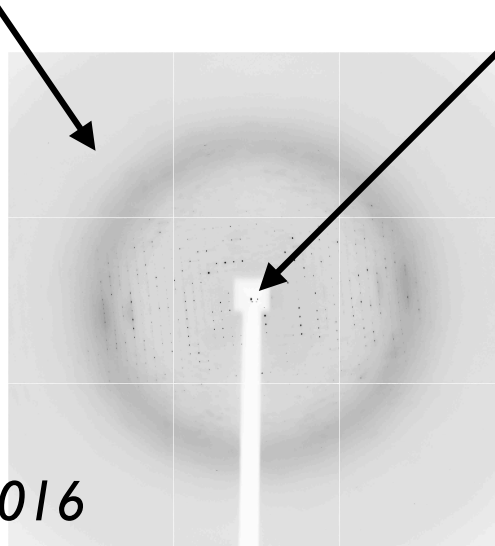
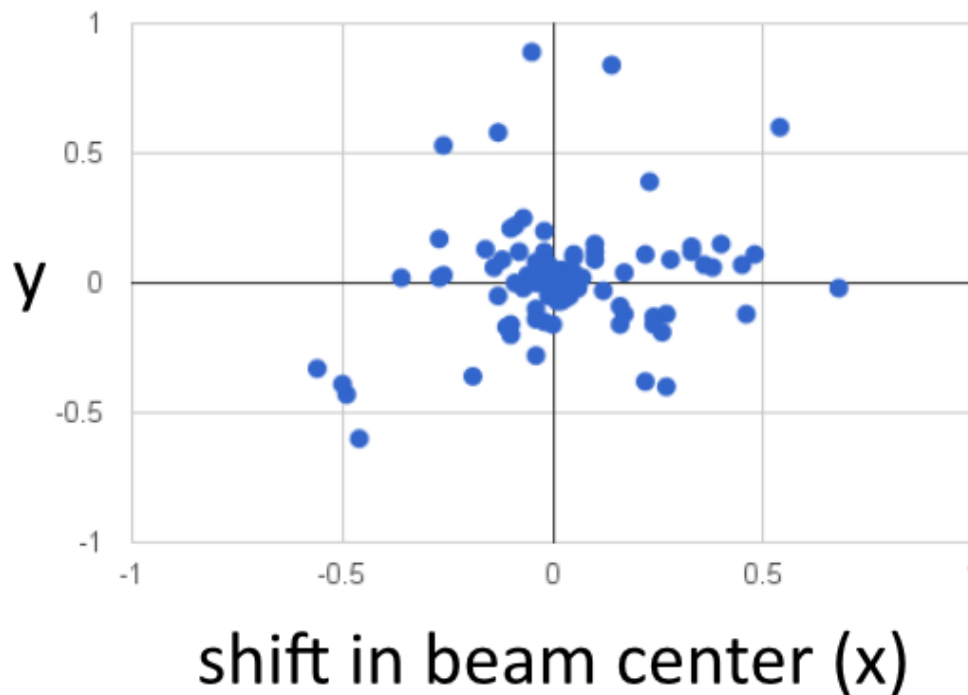
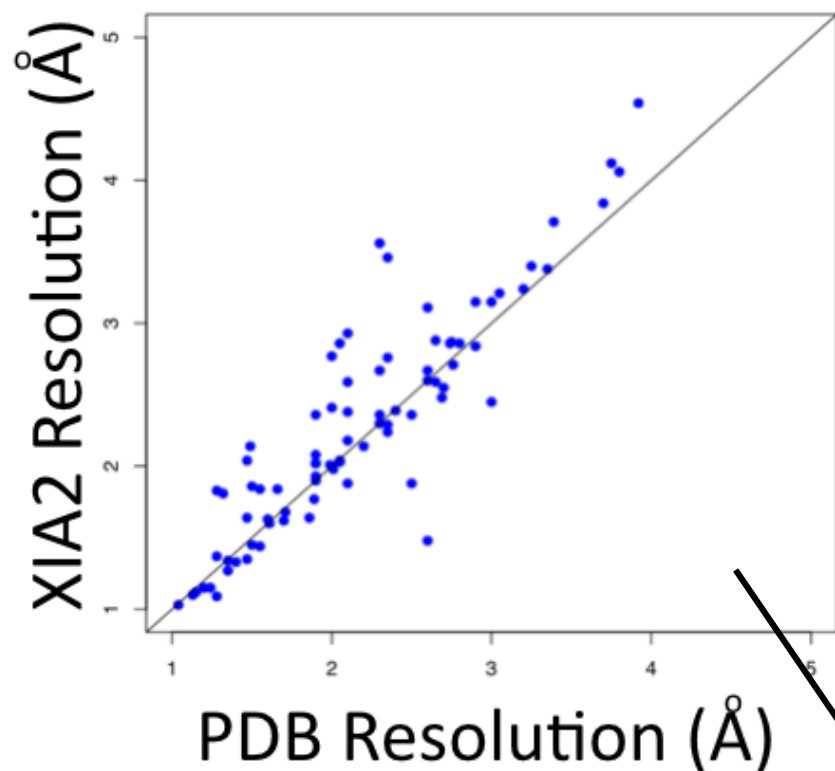
Datasets from the Harvard Medical School

- X-Ray Diffraction data from Brd4 in complex with compound 32, source of structure**
Native dataset
Data DOI: [10.15785/SBGRID/184](https://doi.org/10.15785/SBGRID/184) | Publication DOI: [10.1021/jm501120z](https://doi.org/10.1021/jm501120z)
Blacklow Laboratory, Harvard Medical School
- X-Ray Diffraction data from M2 muscarinic acetylcholine receptor, source of 4MQT structure**
This dataset is compiled from 18 crystals of M2 receptor grown in the presence of the agonist iperexo and the allosteric modulator LY2119620.
Data DOI: [10.15785/SBGRID/125](https://doi.org/10.15785/SBGRID/125) | PDB ID: [4MQT](https://doi.org/10.1038/nature12735) | Publication DOI: [10.1038/nature12735](https://doi.org/10.1038/nature12735)
Kruse Laboratory, Harvard Medical School
- X-Ray Diffraction data from SpyTag/SpyCatcher, source of 4MLI structure**
Native data
Data DOI: [10.15785/SBGRID/90](https://doi.org/10.15785/SBGRID/90) | PDB ID: [4MLI](https://doi.org/10.1016/j.jmb.2013.10.021) | Publication DOI: [10.1016/j.jmb.2013.10.021](https://doi.org/10.1016/j.jmb.2013.10.021)
Rapoport Laboratory, Harvard Medical School



Meyer et al, Nature Communications, 2016

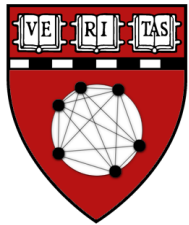
Live-analysis of X-ray diffraction datasets can inject new data analysis methods.



Meyer et al, Nature Communications, 2016

Live Analysis

**The National
DATA SERVICE**



**SBGrid
Consortium**

Phase I (2015)

Phase 2 (2016-2018)

Prototype

**Dataverse
Implementation**



[Root Dataverse](#) > [Shamoo Laboratory Dataverse](#) >

X-Ray Diffraction data from Structures of the E.faecalis LiaR DNA binding domain complexed to the putative consensus sequence, source of 4WU4 structure

Metrics

0 Downloads



X-Ray Diffraction data from Structures of the E.faecalis LiaR DNA binding domain complexed to the putative consensus sequence, source of 4WU4 structure

Shamoo, Yousif, 2016, "X-Ray Diffraction data from Structures of the E.faecalis LiaR DNA binding domain complexed to the putative consensus sequence, source of 4WU4 structure", doi:10.15785/SBGRID/71, Root Dataverse, V1

Cite Data ▾

Learn about [Data Citation Standards](#).

Related Publication

Davlieva M, Shi Y, Leonard PG, et al. A variable DNA recognition site organization establishes the LiaR-mediated cell envelope stress response of enterococci to daptomycin. *Nucleic Acids Research* 2015; 43:4758–4773. doi: 10.1093/nar/gkv321

4WU4 Coordinates

[PDB](#), [MMDB](#)

Biological Sample

Structures of the E.faecalis LiaR DNA binding domain complexed to the putative consensus sequence

Dataset Type

X-Ray Diffraction

Subject Composition

DNA

Data Creation Date

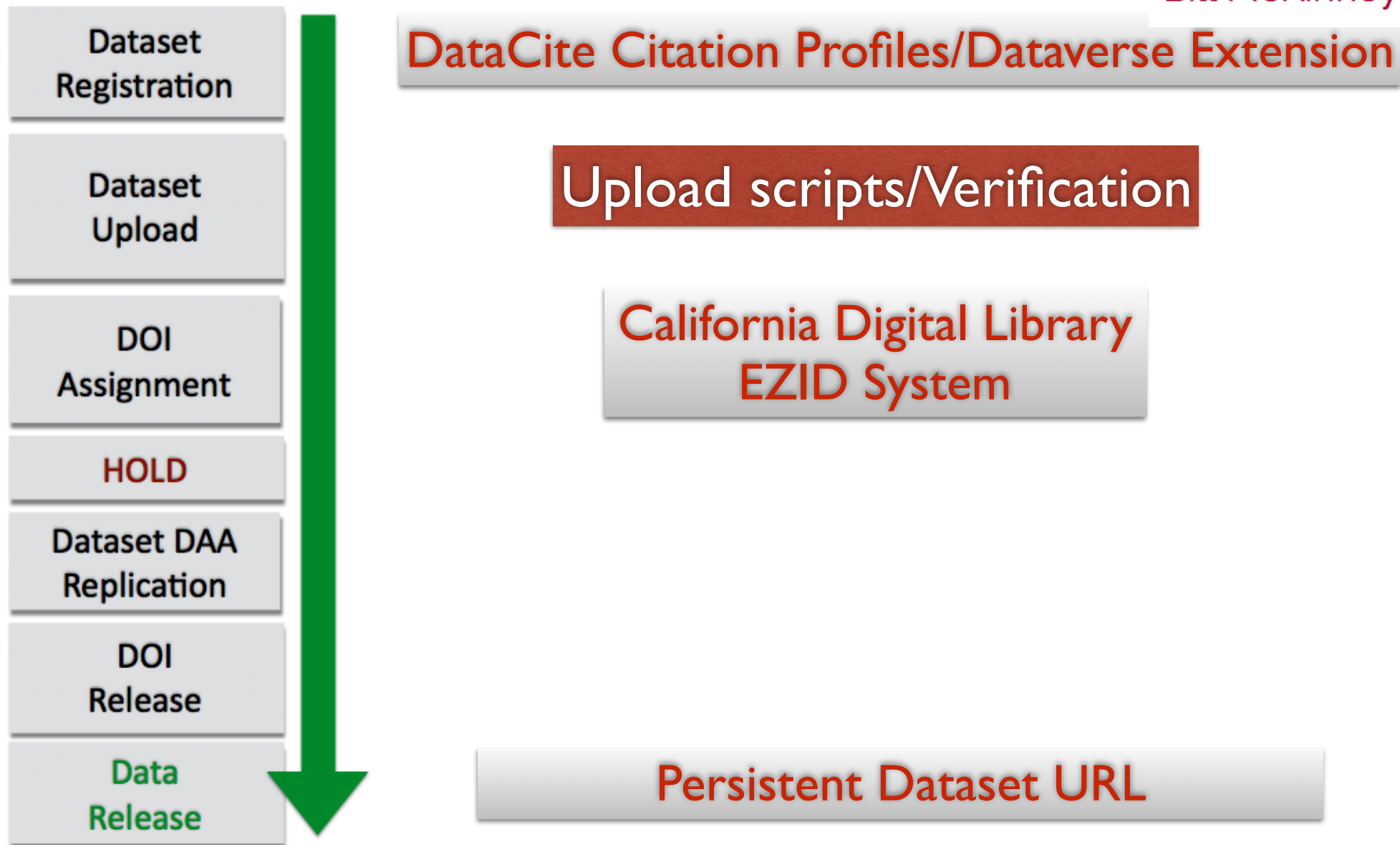
2014-04-19

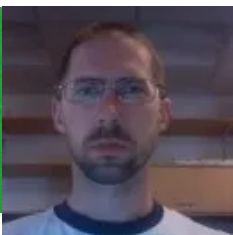
[Visualization](#)
[Files](#)
[Metadata](#)
[Terms](#)
[Versions](#)

Registration followed by Data Transfer Step



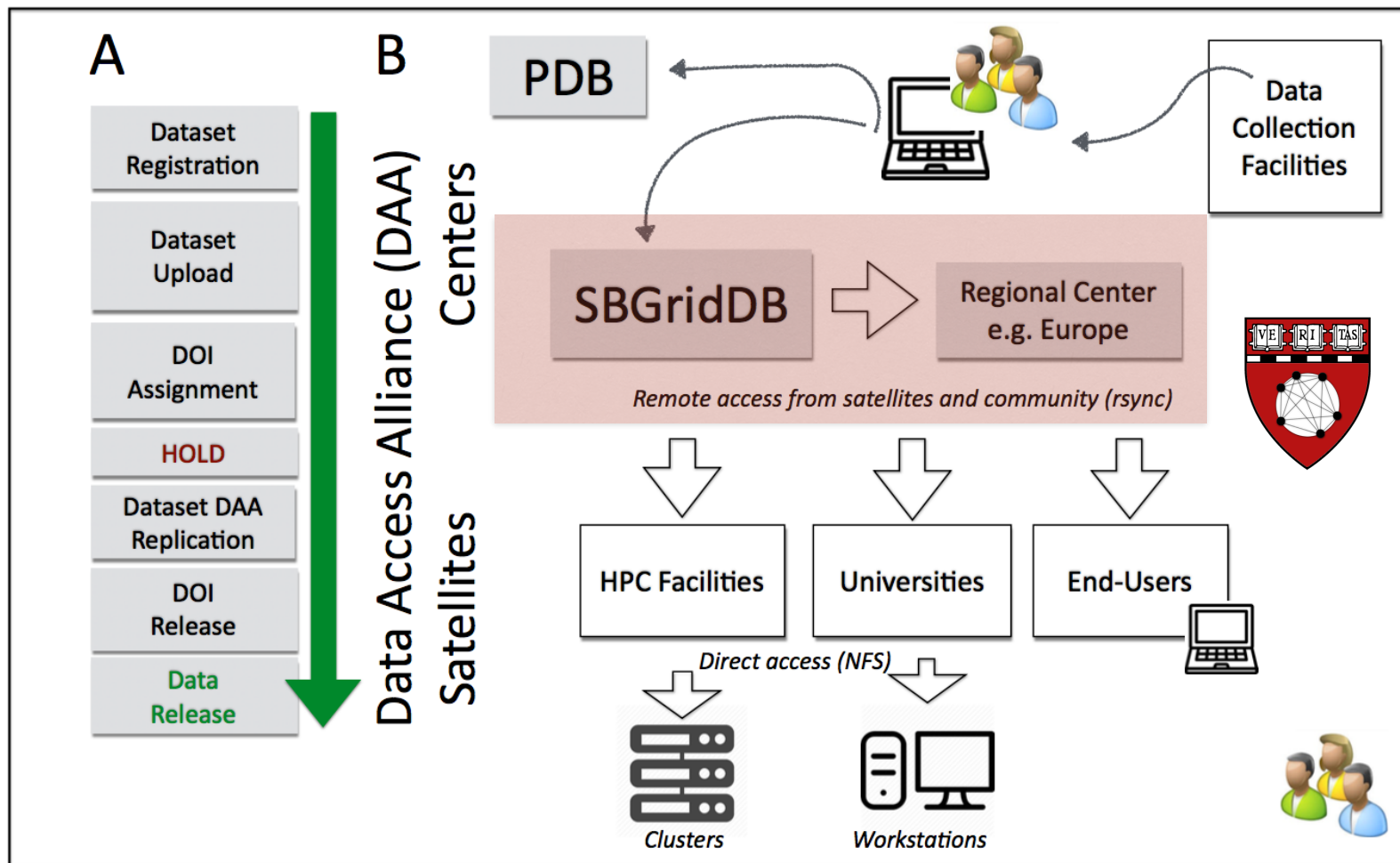
Bill McKinney





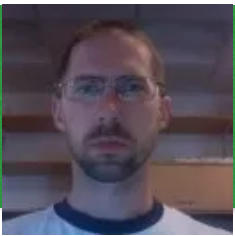
Dataset Access Alliance: Regional access through DAA Centers

Pete
Meyer



Data Access

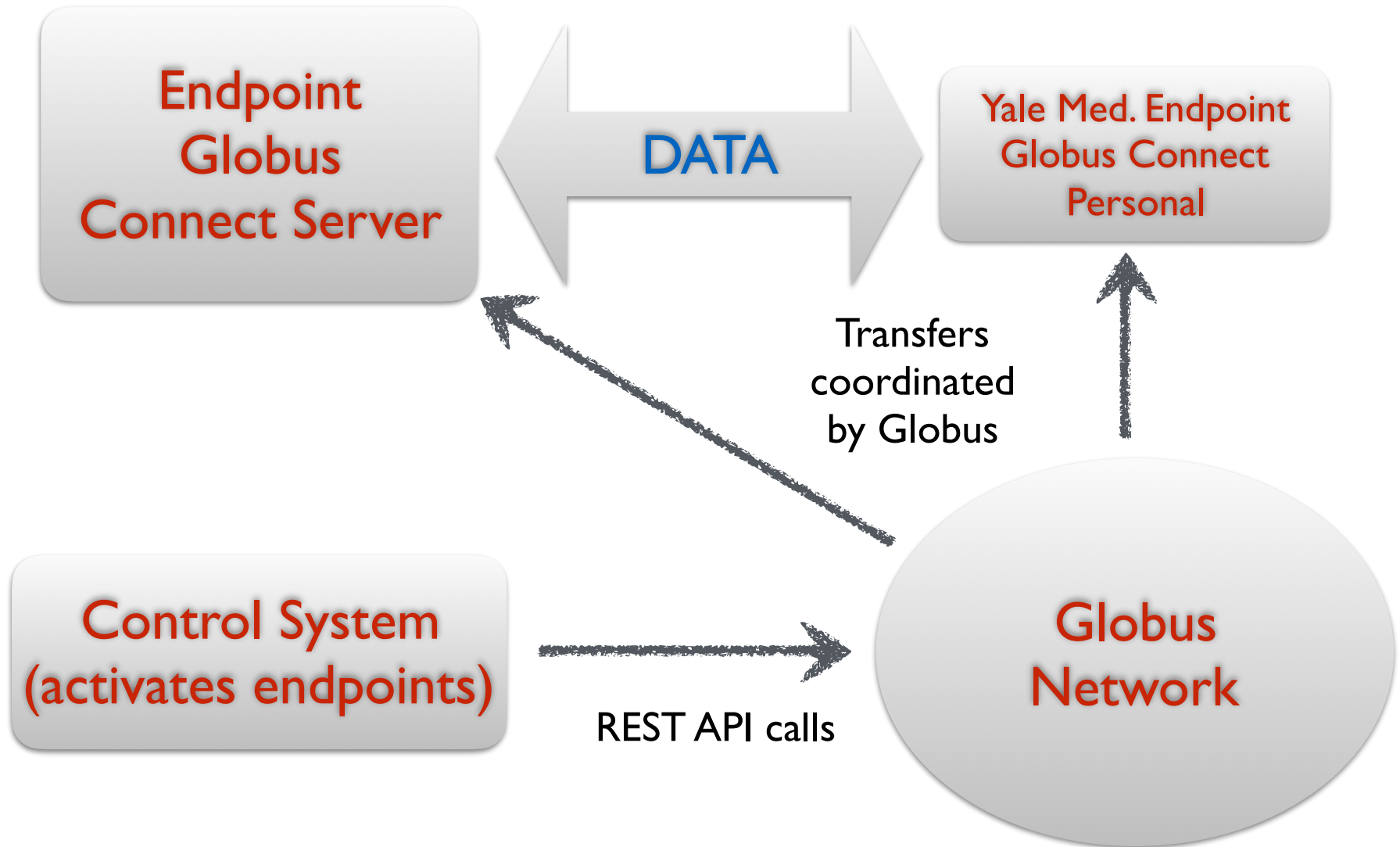
Meyer et al, Nature Communications, 2016



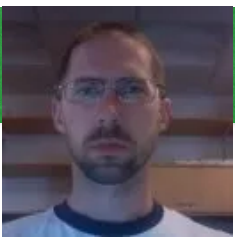
Pete Meyer

Data Access Alliance: Globus Infrastructure

**The National
DATA SERVICE**



Data Locality Module



Pete Meyer



Dataverse Server



Globus Control System

Data Access Alliance Sites

XD 1

XD 2

EM

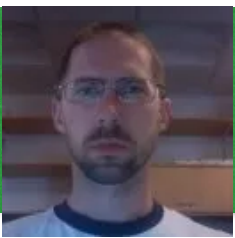
LLSM

Institution 1

Institution 2

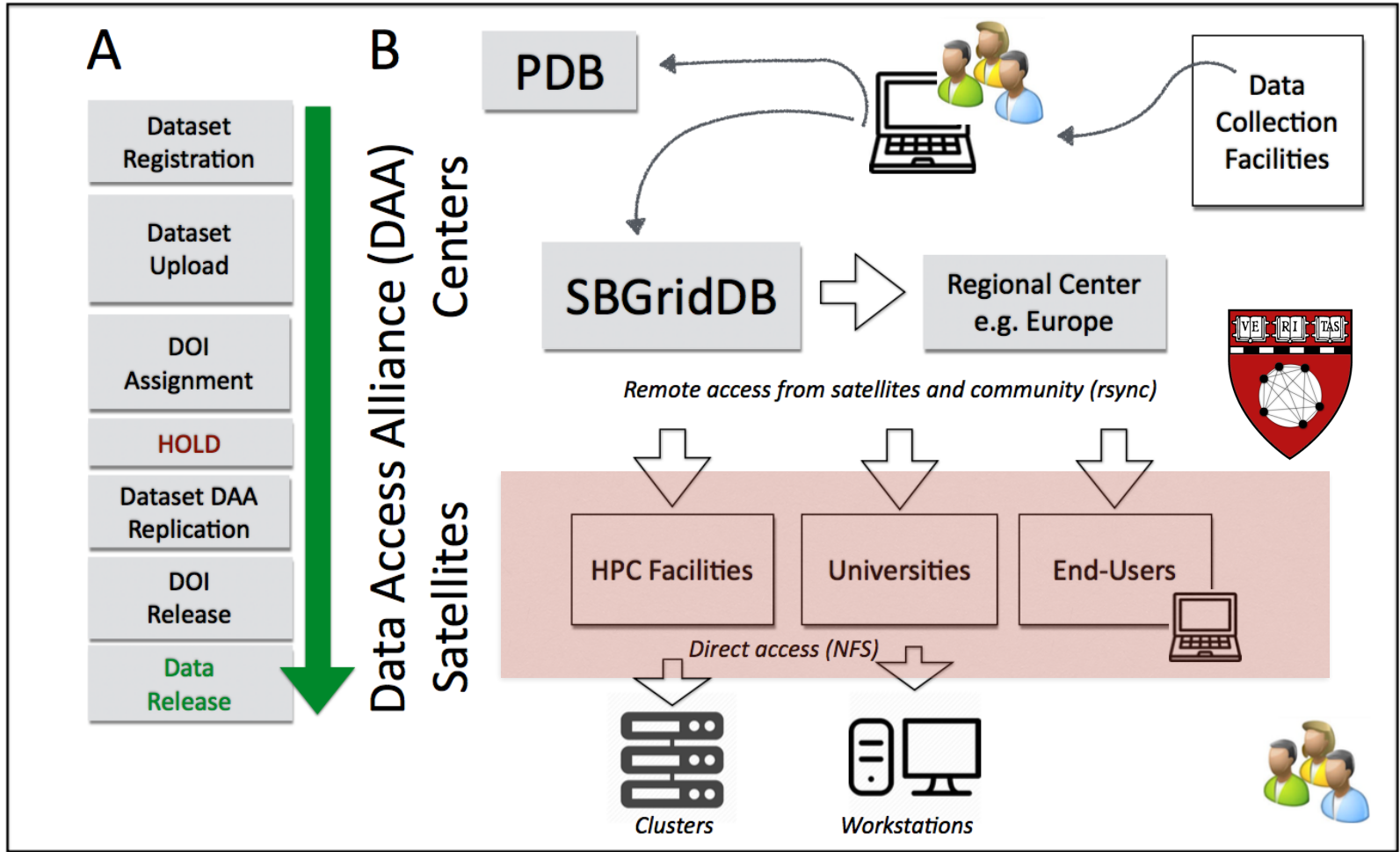


- reporting new datasets to other systems for replication
- user initiated replication
 - ▶ retention policies
- admin initiated replication



Pete Meyer

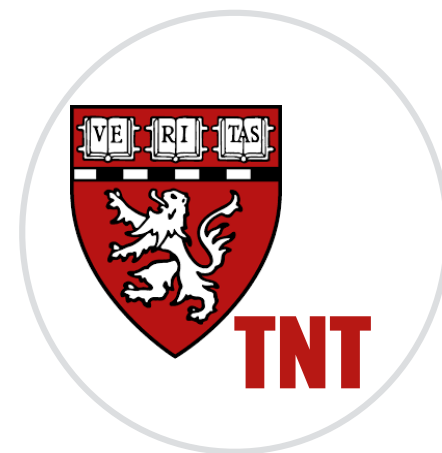
Dataset Access Alliance: Local access through SBGrid Satellites



SBx: Desktop Tool for Structural Biologists



Jason Key



Apply changes Refresh

Packages

- Installed
- Pending

Collections

- All (51)
- Crystallography (4)
- NMR (4)
- Electron Microscopy (9)
- Structure Visualization & Analysis (6)
- Computational Chemistry (6)
- Other (22)

Selected	Package	Installed Version	Current Version
<input type="checkbox"/>	a2ps		4.14
<input type="checkbox"/>	Aline		1.0.025
<input type="checkbox"/>	AMPS		2.3a
<input type="checkbox"/>	ARIA		2.3.1
<input type="checkbox"/>	AutoDock		4.2.5.1
<input type="checkbox"/>	BLAST		2.2.26
<input type="checkbox"/>	BLAST+		2.2.31
<input type="checkbox"/>	Bowtie		1.0.0
<input type="checkbox"/>	Bowtie 2		2.2.1
<input type="checkbox"/>	breseq		1.00rc8
<input type="checkbox"/>	BWA		0.7.7-r441
<input type="checkbox"/>	CCP4		6.5
<input type="checkbox"/>	Chimera		1.10.2
<input type="checkbox"/>	CNS		1.3r8
<input type="checkbox"/>	CTF		20140609
<input checked="" type="checkbox"/>	Cufflinks		2.1.1
<input type="checkbox"/>	cython		0.21.1
<input type="checkbox"/>	DireX		0.7.0
<input type="checkbox"/>	EMAN		1.9
<input type="checkbox"/>	EMAN2		2.12
<input type="checkbox"/>	EPMR		15.04
<input type="checkbox"/>	FASTA		36.3.8a
<input type="checkbox"/>	FREALIGN		9.11_151013
<input type="checkbox"/>	GROMACS		5.1
<input type="checkbox"/>	HADDOCK		2.1
<input type="checkbox"/>	IGV		2.3.34
<input type="checkbox"/>	MAFFT		7.245
<input type="checkbox"/>	MGLTools		1.5.7rc1

Cufflinks

Description

a reference-guided assembler that assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.

Links

- [Website](#)
- [Manual](#)
- [Forum help](#)



Protein Viewer



Rachel
Partridge

Visualization

Files

Metadata

Terms

Versions

Choose Style:

Preset

Cartoon

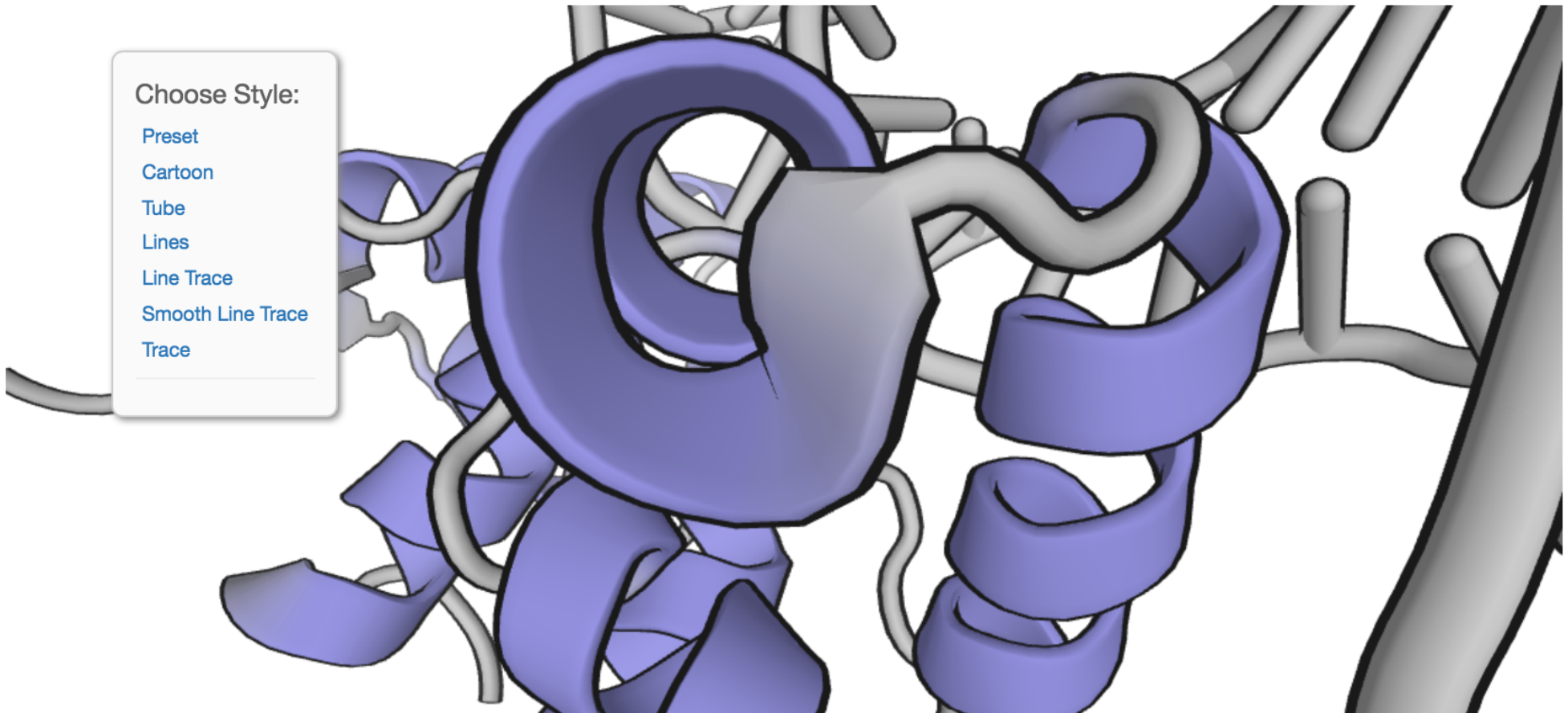
Tube

Lines

Line Trace

Smooth Line Trace

Trace



Community-wide adoption

[Home](#) [NSF Workshop](#) [Journal-Repository Workshop](#) [Register](#) [Travel/Lodging](#)

[#DSCBOSTON2016](#) help@sbgrid.org

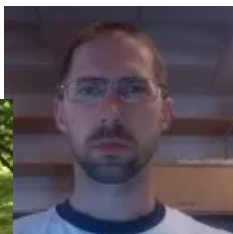


Connecting Journals
to Data Repositories

June 7, Harvard Medical School, Boston, MA



Stephanie Socias
Meyer



Pete Meyer



Bill McKinney



Rachel Partridge



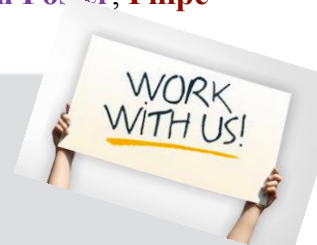
Mercè Crosas

**Phil Durbin,
Eleni Castro,
Gustavo Durand,
Leonid Andreev**

Dataverse Team: Gustavo Durand, Leonid Andreev, Stephen Kraffmiller, Phil Durbin, Raman Prasad, Eleni Castro, Danny Brooke

Pete Meyer, Stephanie Socias, Jason Key, Elizabeth Ransey, Emily C. Tjon, Alejandro Buschiazzo, Ming Lei, Chris Botka, James Withrow, David Neu, Kanagalaghatta Rajashankar, Karen S. Anderson, Richard Baxter, Stephen Blacklow, Titus J. Boggon, Alexandre M.J.J. Bonvin, Dominika Borek, Tom J. Brett, Amedeo Caflisch, Chung-I Chang, Walter J. Chazin, Kevin D. Corbett, Michael S. Cosgrove, Sean Crosson, Sirano Dhe-Paganon, Enrico Di Cera, Catherine L. Drennan, Michael J. Eck, Brandt F. Eichman, Qing R. Fan, Adrian R. Ferré-D'Amaré, James S. Fraser, J. Christopher Fromme, K. Christopher Garcia, Rachelle Gaudet, Peng Gong, Stephen Harrison, Ekaterina E. Heldwein, Zongchao Jia, Robert J. Keenan, Andrew C. Kruse, Marc Kvangsakul, Jason S. McLellan, Yorgo Modis, Yunsun Nam, Zbyszek Otwinowski, Emil F. Pai, Pedro José Barbosa Pereira, Carlo Petosa, CS Raman, Tom A. Rapoport, Antonina Roll-Mecak, Michael K. Rosen, Gabby Rudenko, Joseph Schlessinger, Thomas U. Schwartz, Yousif Shamoo, Holger Sondermann, Yizhi J. Tao, Niraj H. Tolia, Oleg V. Tsodikov, Kenneth D. Westover, Hao Wu, Ian Foster, Filipe Maia, Tamir Gonen Tom Kirchhausen, Merce Crosas, Piotr Sliz

- The Leona M. and Harry B. Helmsley Charitable Trust 2016PG-BRI002 to PS and MC.
- NSF SI2 1448069 (to P.S.)
- NCRRI 1S10RR028832 (HMS)
- NIH, NIH Intramural Program, HHMI, EU Infrastructure Grant, The Swiss National Science Foundation, National Science and Engineering Research Council of Canada, McKnight Scholar Award, Wellcome Trust, Canadian Institutes of Health, ANRS/Fondation de France, Fundação para a Ciência e a Tecnologia, Portugal, Welch Foundation, Edward Mallinckrodt, Jr. Foundation, CPRIT.



Meyer et al, Nature Communications, 2016