THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

The Dataverse Project

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

Boston Children's Hospital

Until every child is well™

Harvard Medical School

July 11th, 2016

Dataverse Community Meeting 2016

Dataverse Repositories Around the World

Sonia Barbosa

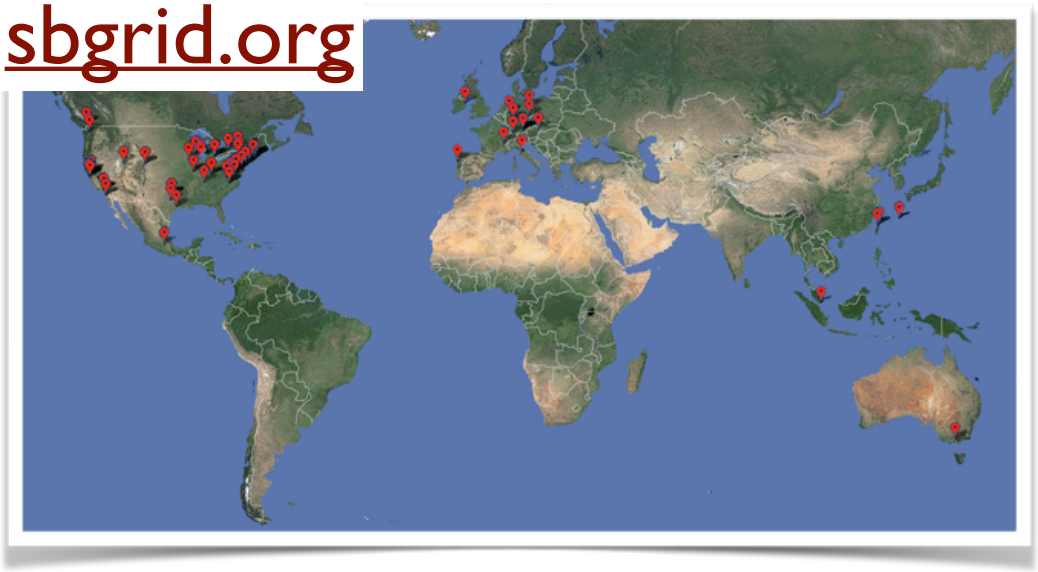
2000-2016: Growth of the community

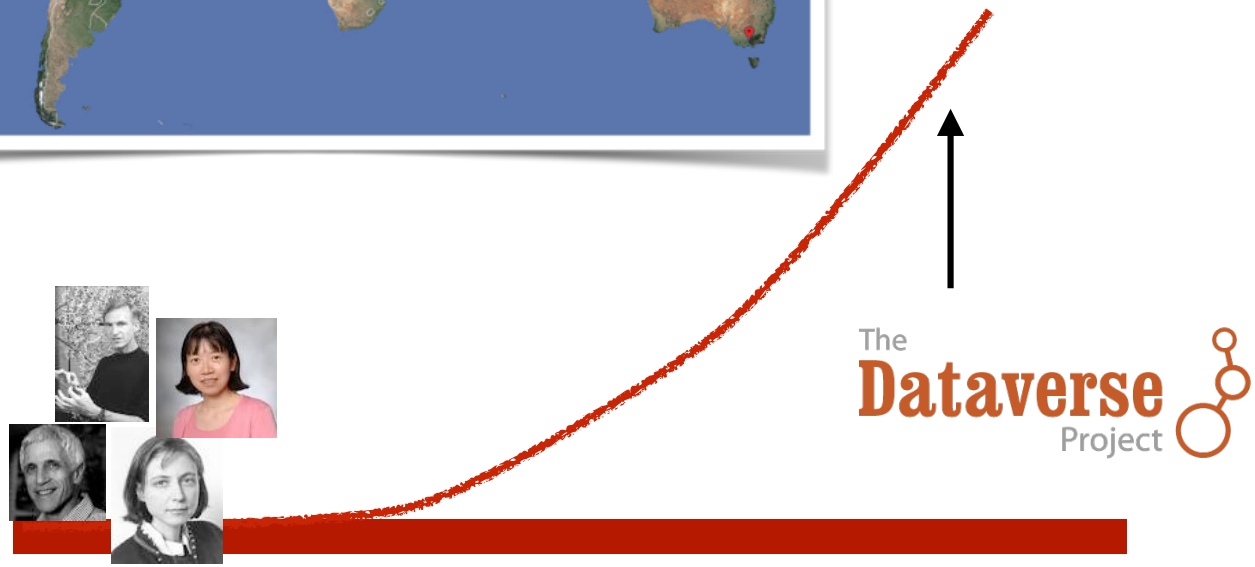
~300 structural biology groups

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

22 countries

sbgrid.org





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

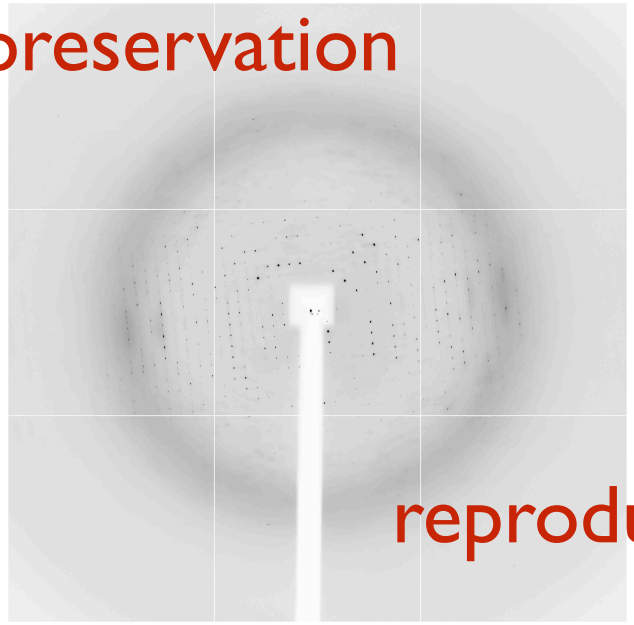
- Software Support
  - (~300 OS, academic,commercial applications)
- Local Research Computing Infrastructure
  - Structural Biology (clusters, storage, workstations)
  - Lattice Light Sheet Microscopy
- Access to HPC Resources
  - Open Science Grid (Portal)
  - XSEDE
- Teaching and Training
  - YouTube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

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

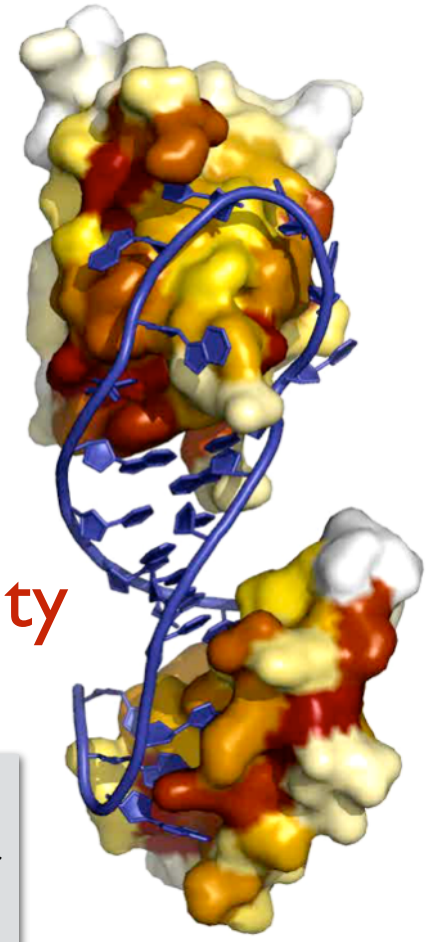
Motivation

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

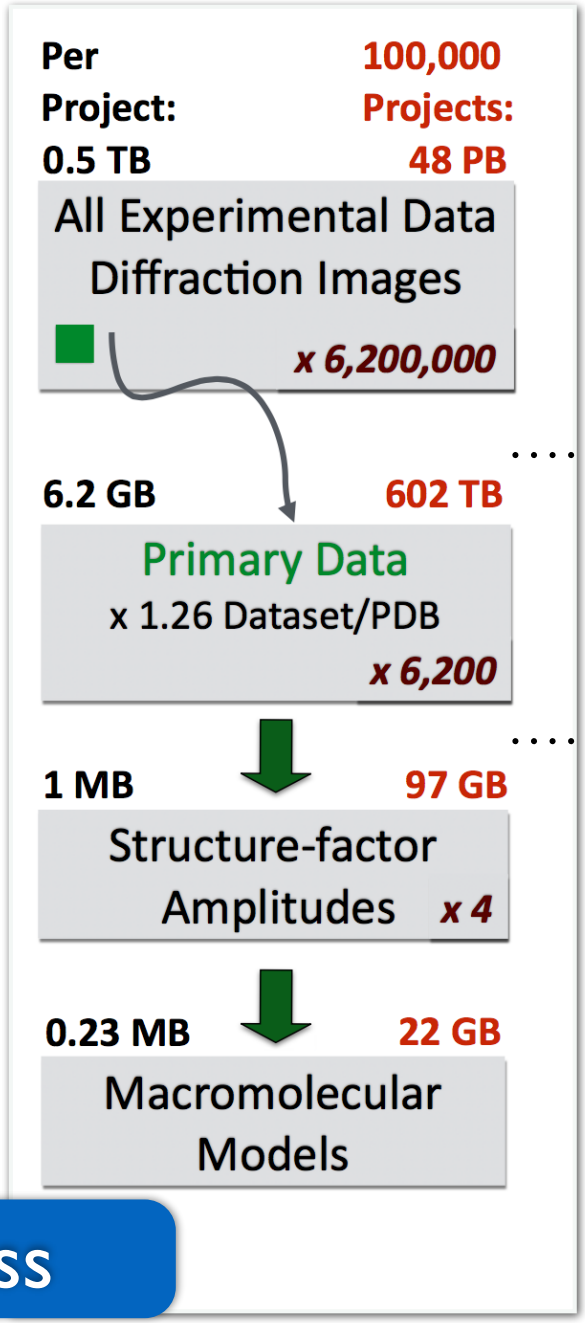
MD Models

- 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 anisotropicity)
- Additional features: (e.g. anisotropic diffuse scattering signals)

WORLDWIDE PDB
PROTEIN DATA BANK

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

SBGrid CONSORTIUM

Molecular models and reduced datasets are stored in Protein Data Bank

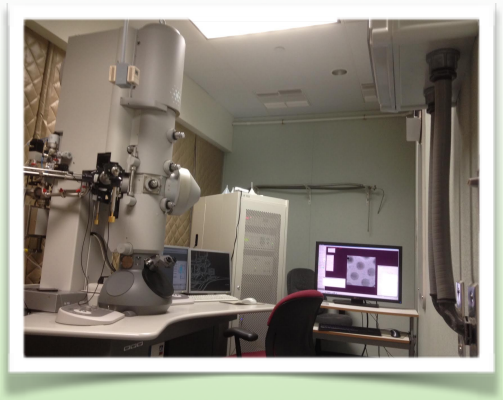
WORLDWIDE PDB PROTEIN DATA BANK

Data Access

Meyer et al, Nature Communications, 2016

Motivation

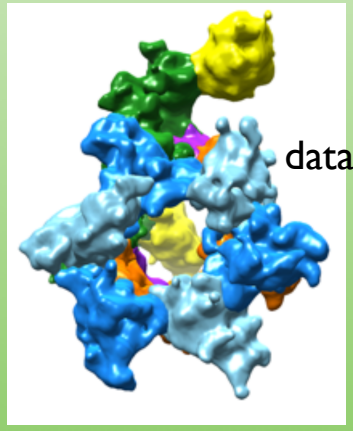
Example II: EM structure - management of datasets from EM facilities



HARVARD MEDICAL SCHOOL

K2 Computer

Drift Correction (GPU)



uncorrected data (~10TB/dataset)

"summed" images

Fast Storage

Local Processing
2x20 Core Xeon 2600 V4s
(20TB scratch)

Orchestra/XSEDE

Local Archives

UMASS

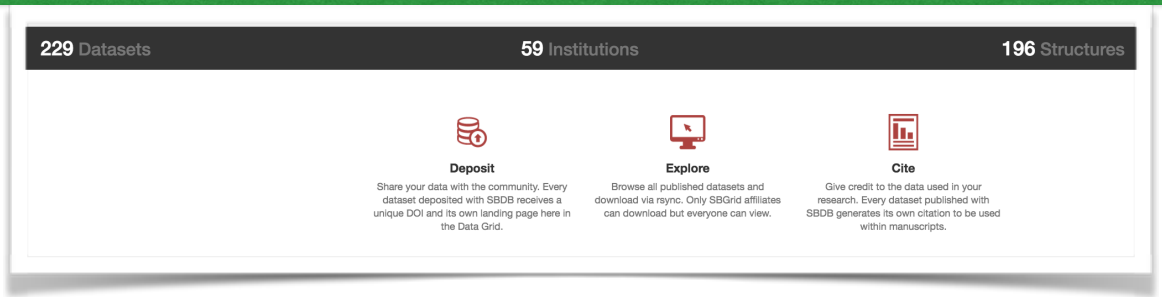
K2 Computer

Titan Krios 300kV TEM
Talos-Arctica 200 kV

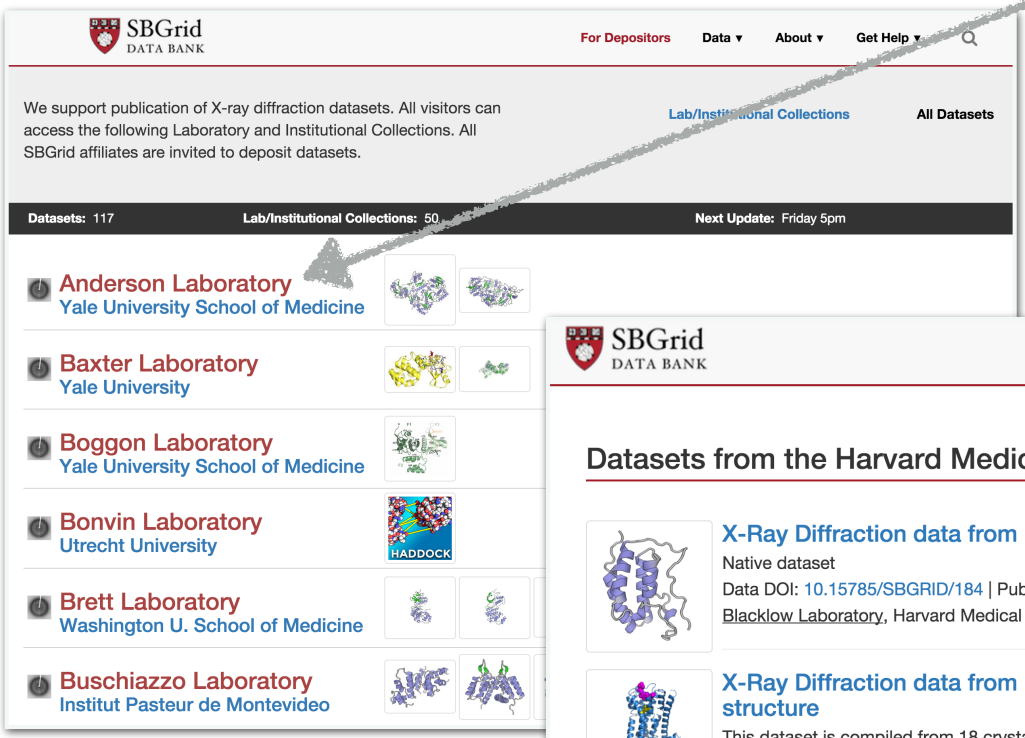
Drift Correction (GPU)

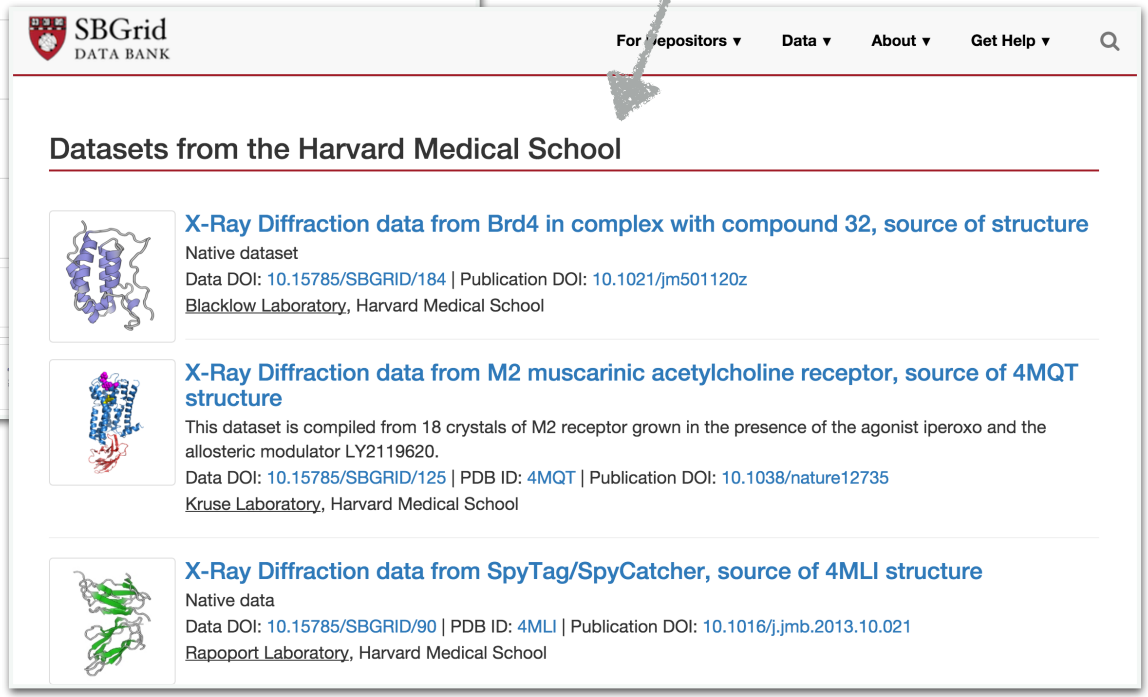
WORLDWIDE PDB
PROTEIN DATA BANK

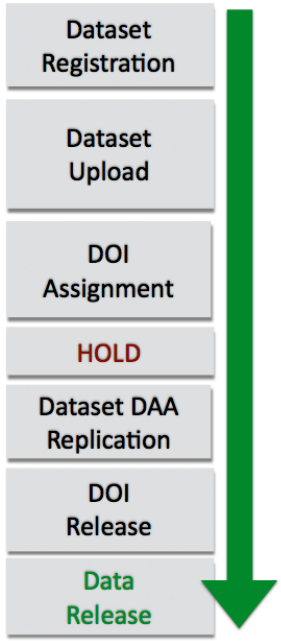
data.sbgrid.org website



Lab Collections Institutional Collections

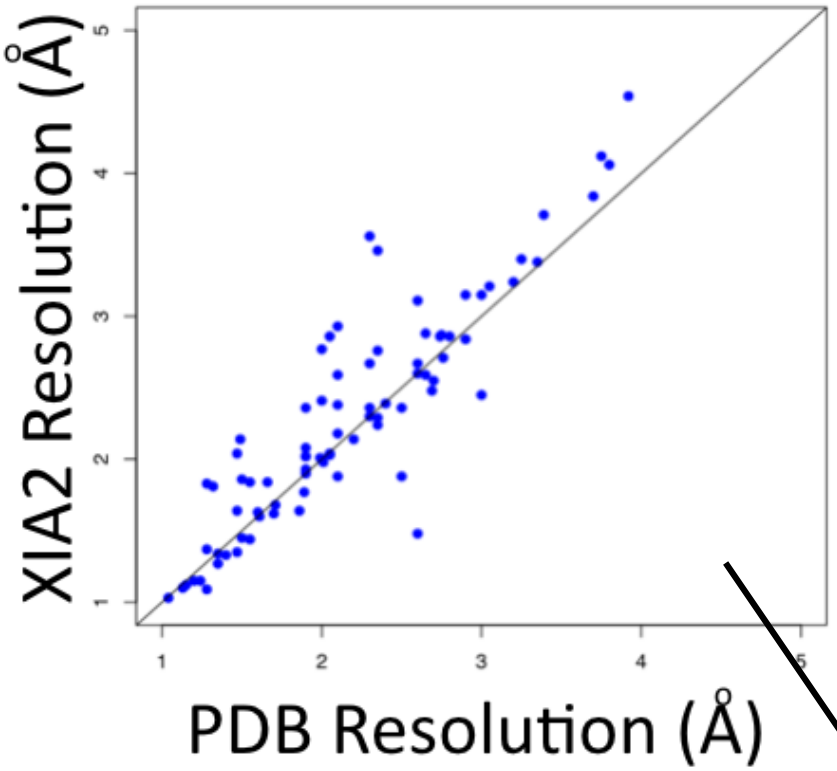


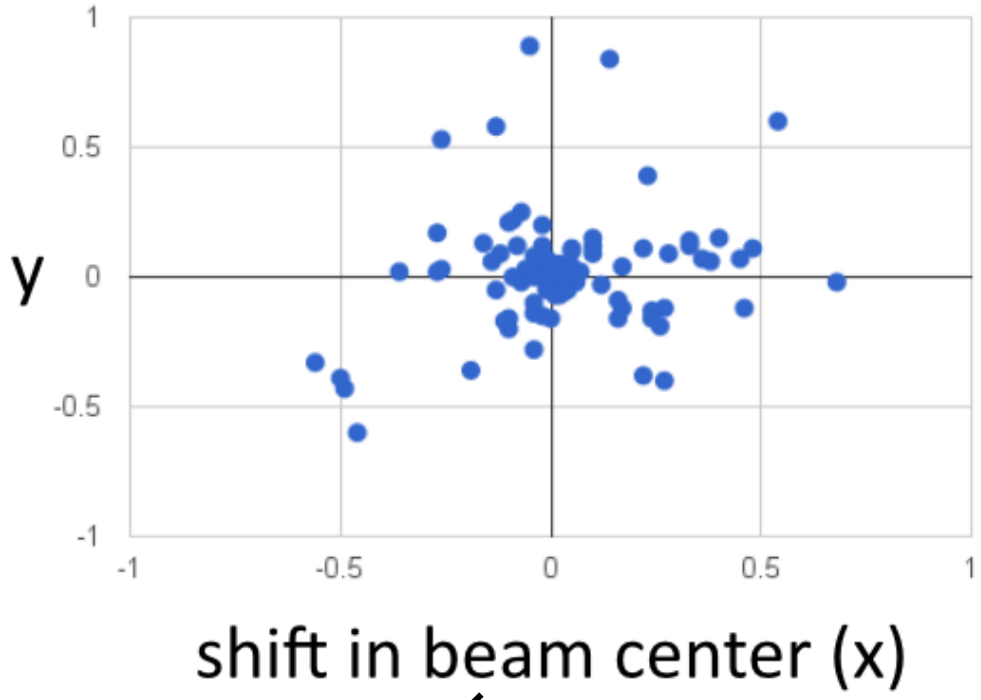


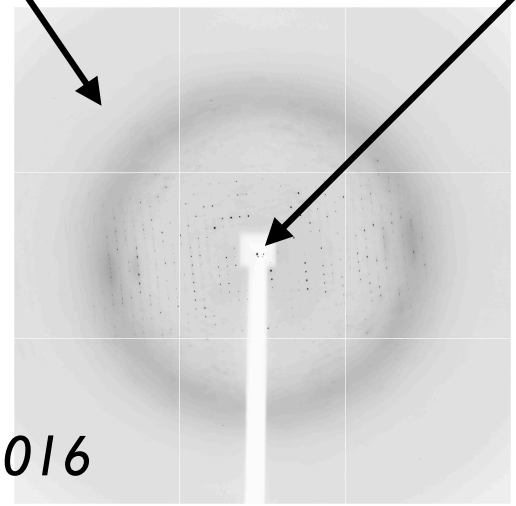


Meyer et al, Nature Communications, 2016

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







Meyer et al, Nature Communications, 2016

Live Analysis

2016

The National DATA SERVICE

globus

SBGrid Consortium

The Dataverse Project

Phase 1 (2015)

Prototype

Phase 2 (2016-2018)

Dataverse Implementation

2016

Bill McKinney

Dataverse Beta: for testing only

About Guides Supp

Shamoo Laboratory Dataverse (Rice University)

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 Registration

Dataset Upload

DOI Assignment

HOLD

Dataset DAA Replication

DOI Release

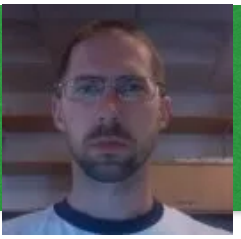
Data Release

DataCite Citation Profiles/Dataverse Extension

Upload scripts/Verification

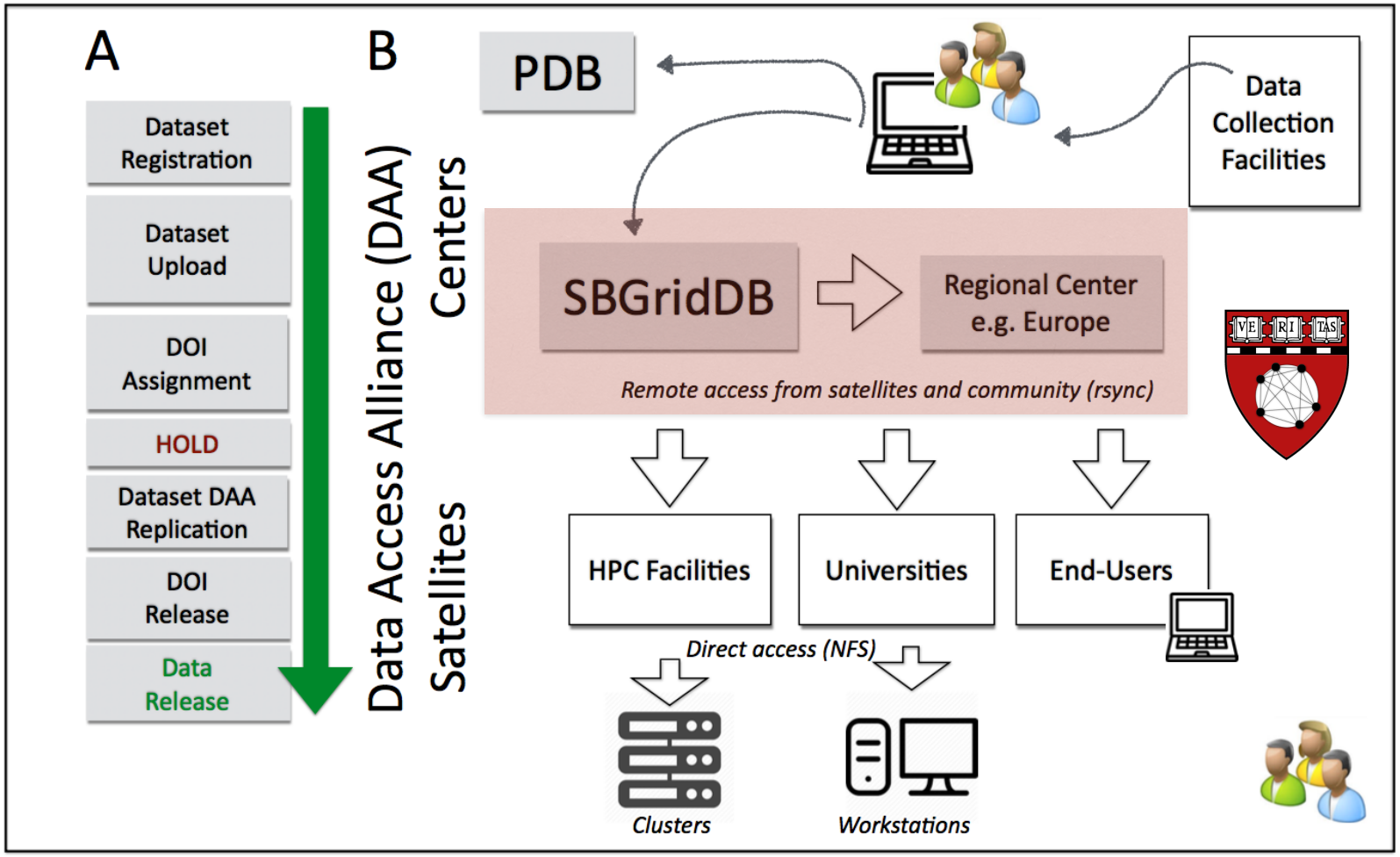
California Digital Library
EZID System

Persistent Dataset URL



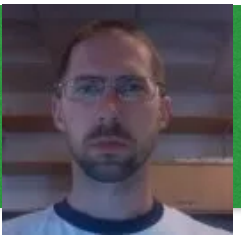
Pete Meyer

Dataset Access Alliance: Regional access through DAA Centers



Data Access

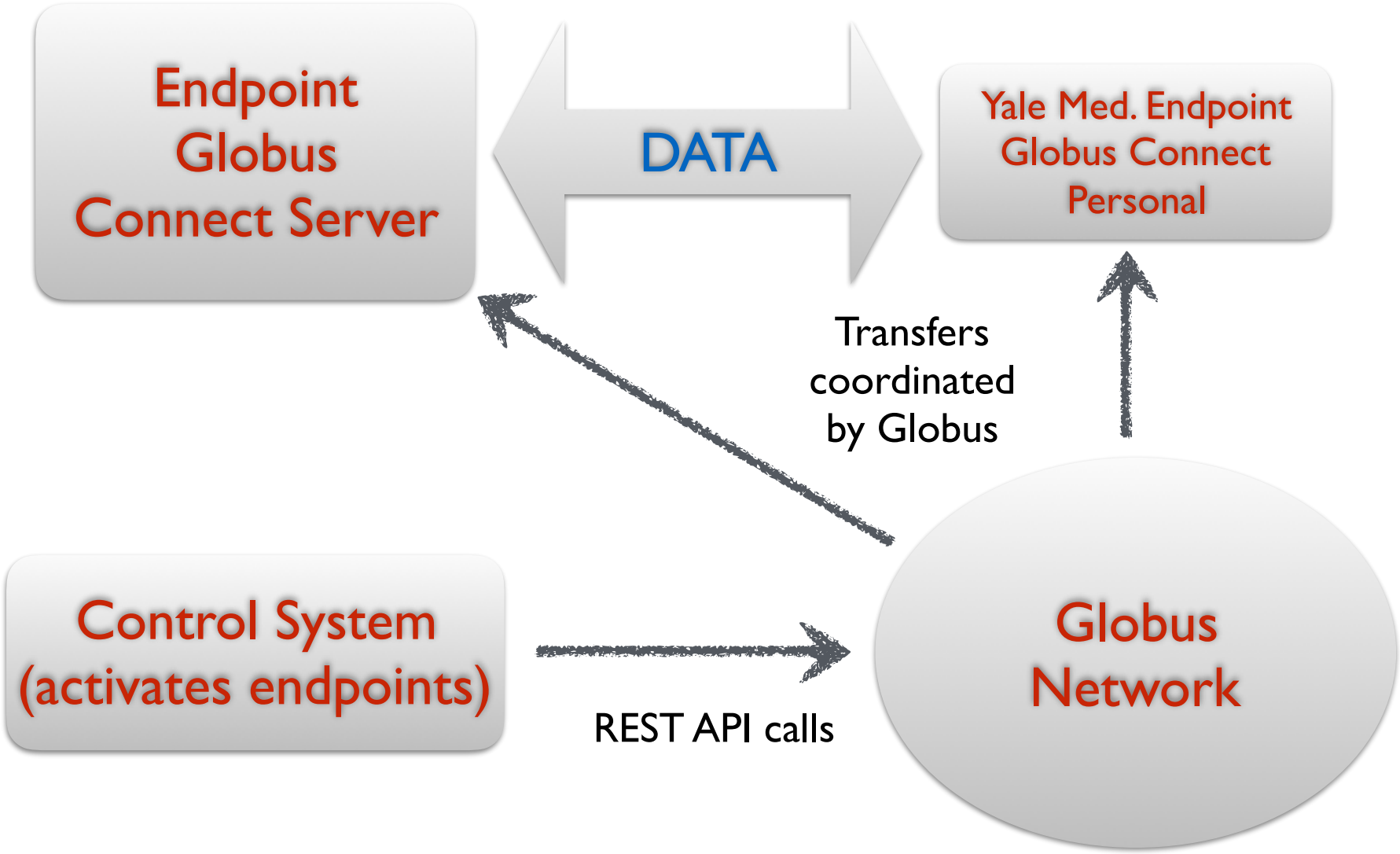
Meyer et al, Nature Communications, 2016



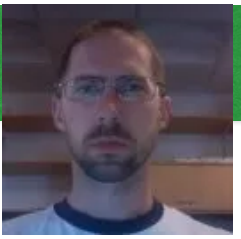
Pete Meyer

Data Access Alliance: Globus Infrastructure

The National DATA SERVICE



Data Locality Module



Pete Meyer

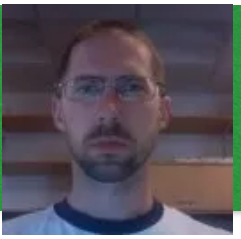
Dataverse Server

Globus Control System

Data Access Alliance Sites

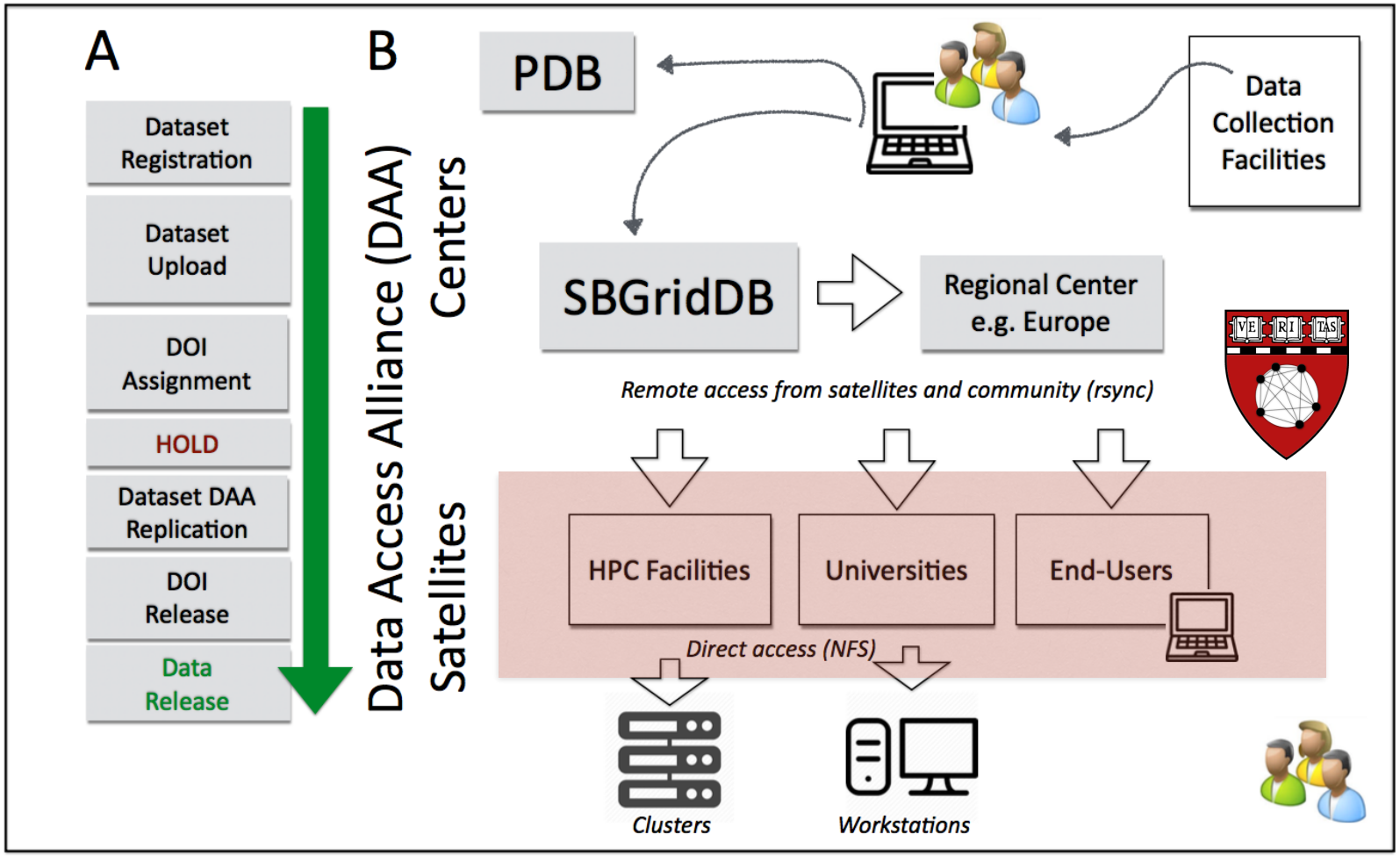
- XD 1
- XD 2
- EM
- LLSM
- Institution 1
- Institution 2

- 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

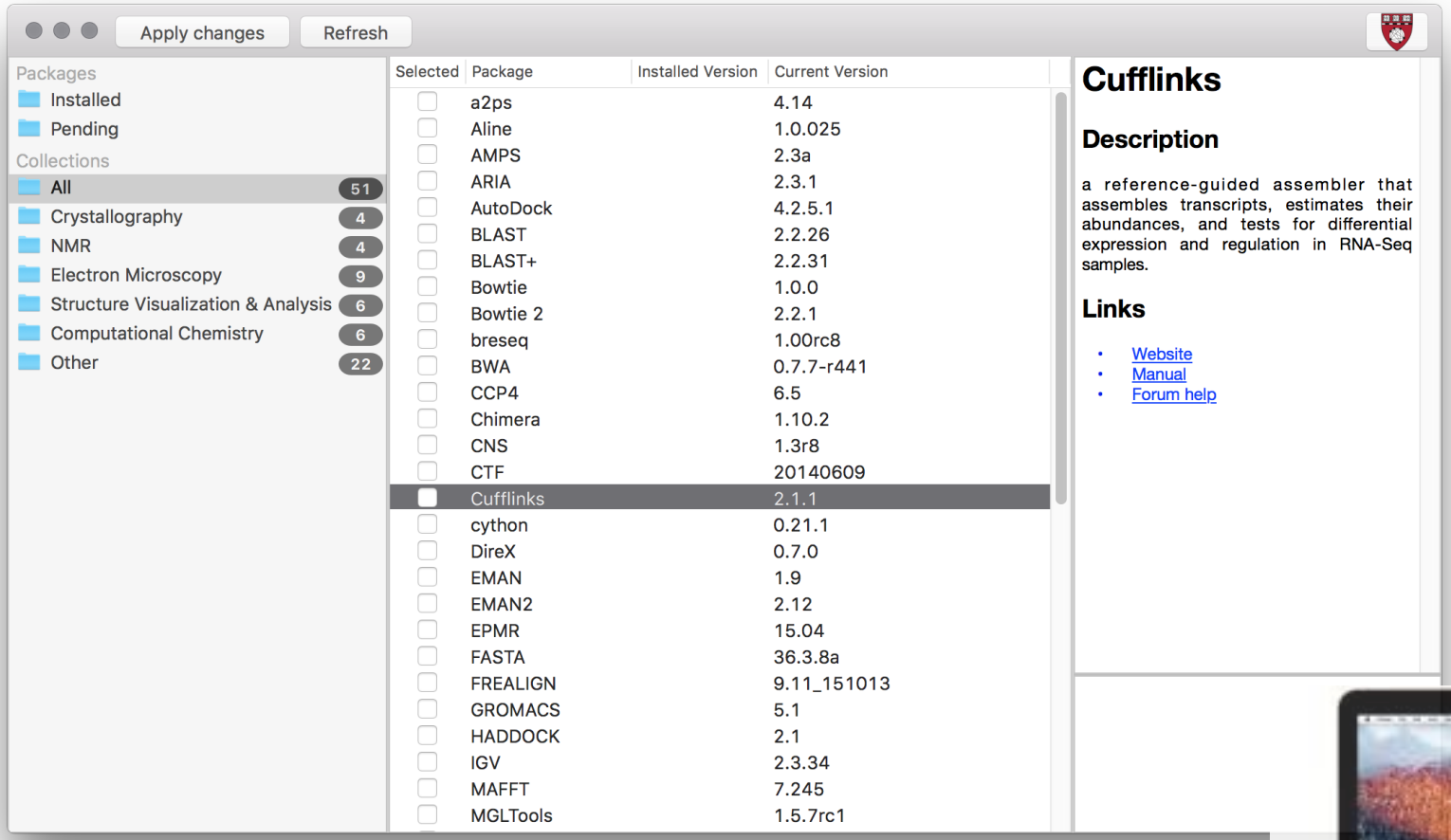


Data Access

Meyer et al, Nature Communications, 2016

SBx: Desktop Tool for Structural Biologists

Jason Key



Protein Viewer

Rachel Partridge

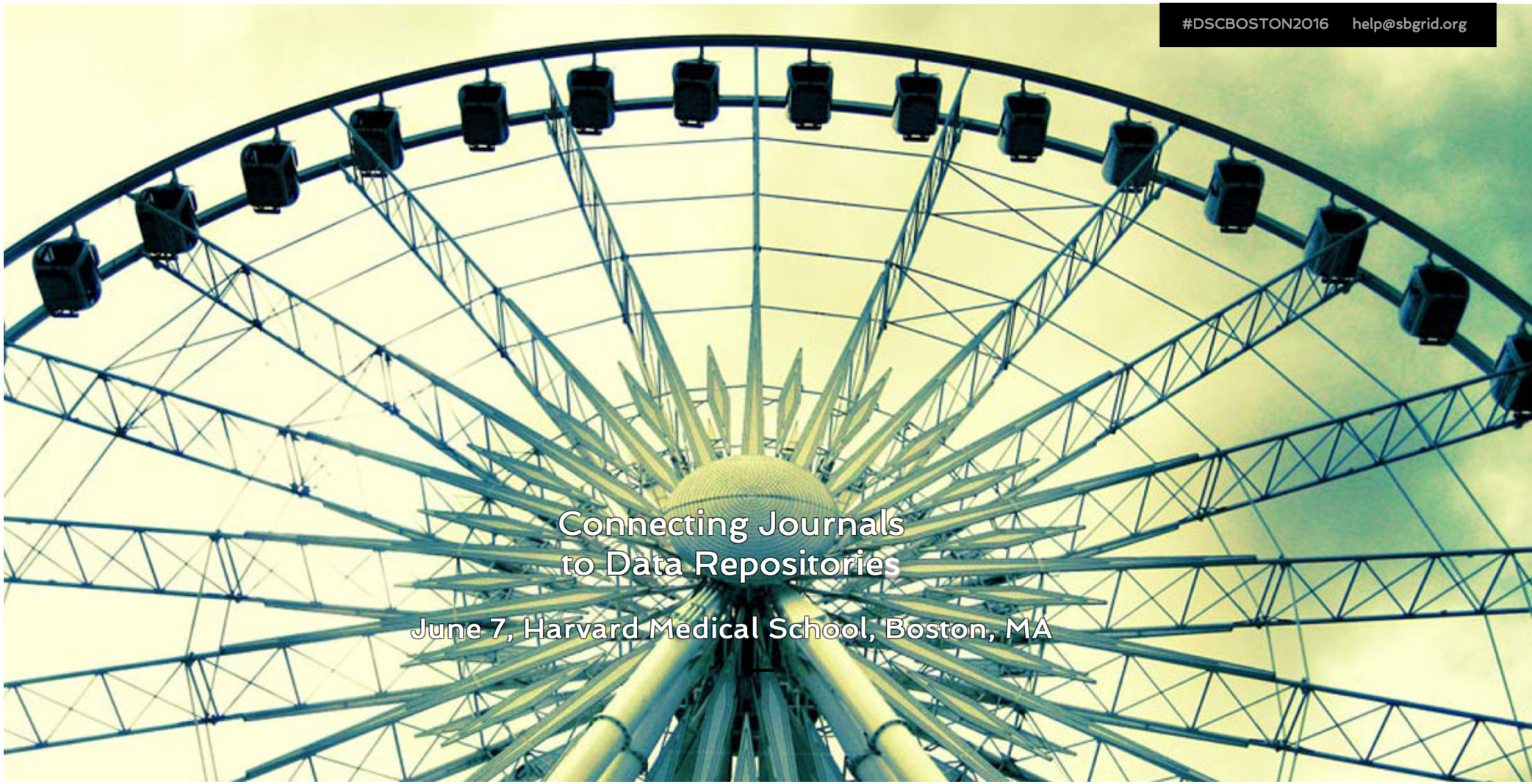
Visualization | Files | Metadata | Terms | Versions

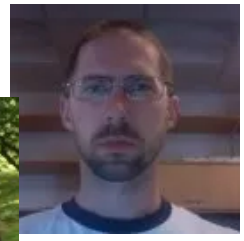
Choose Style:

- Preset
- Cartoon
- Tube
- Lines
- Line Trace
- Smooth Line Trace
- Trace

Community-wide adoption

Home NSF Workshop Journal-Repository Workshop Register Travel/Lodging



Stephanie Socias

Pete Meyer

Bill McKinney

Rachel Partridge

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

Boston Children's Hospital
Until every child is well™

VE RI TAS
Harvard Medical School

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 Neau, 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 Kvansakul, 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.)
- NCRR 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, Welcome 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