THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST

National Data Service Pilot

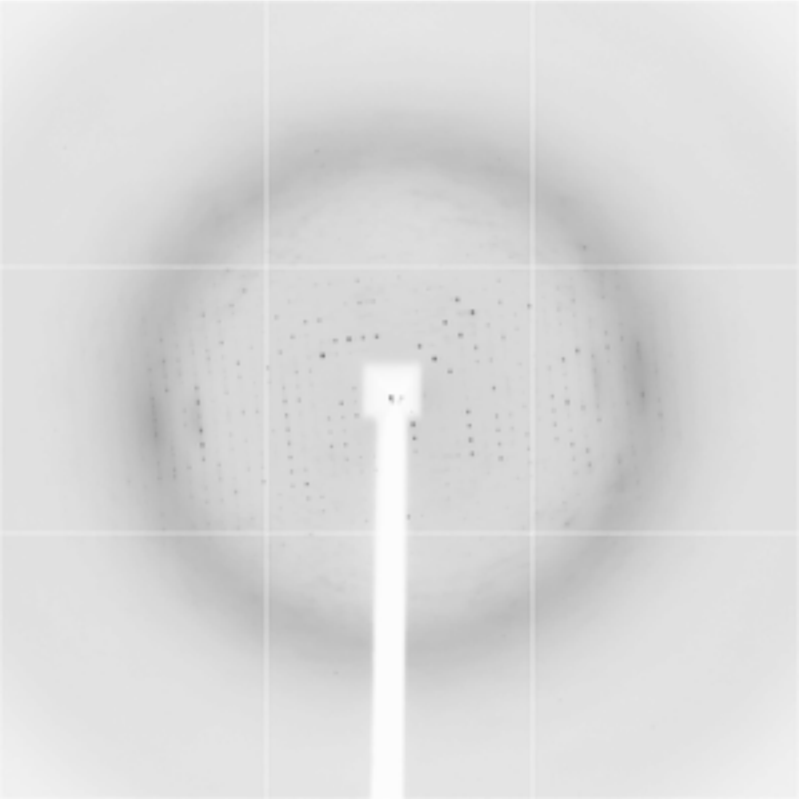
Establishing an effective system to facilitate access to biomedical datasets.

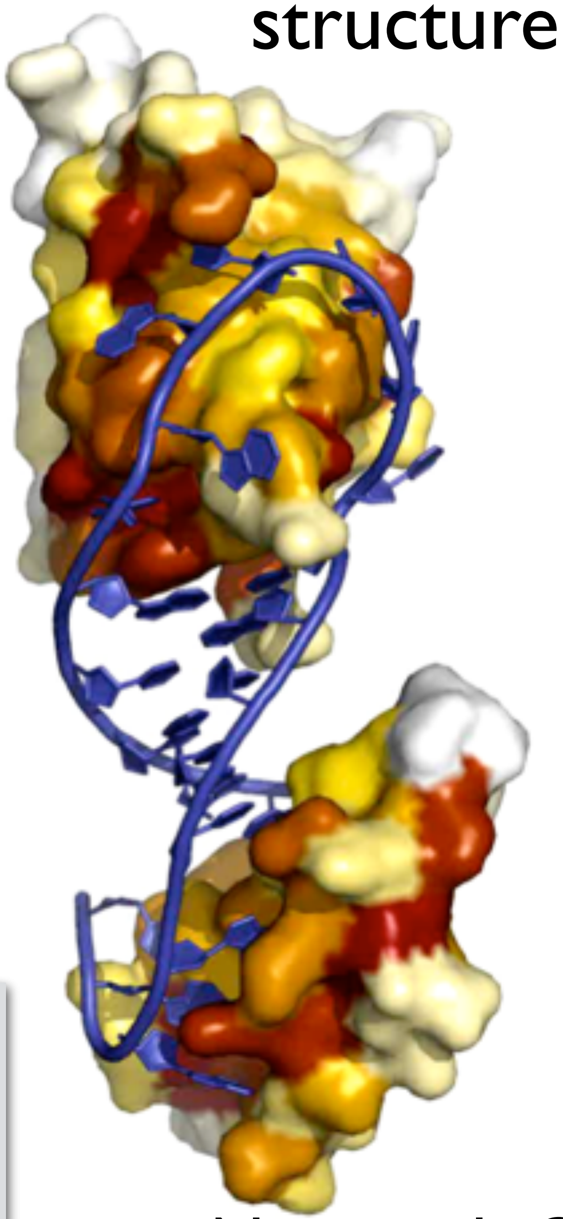
Peter Meyer

Piotr Sliz, Merce Crosas, Ian Foster

National Data Service
UNC-Chapel Hill, April 5th, 2016

Motivation





Lin28 structure

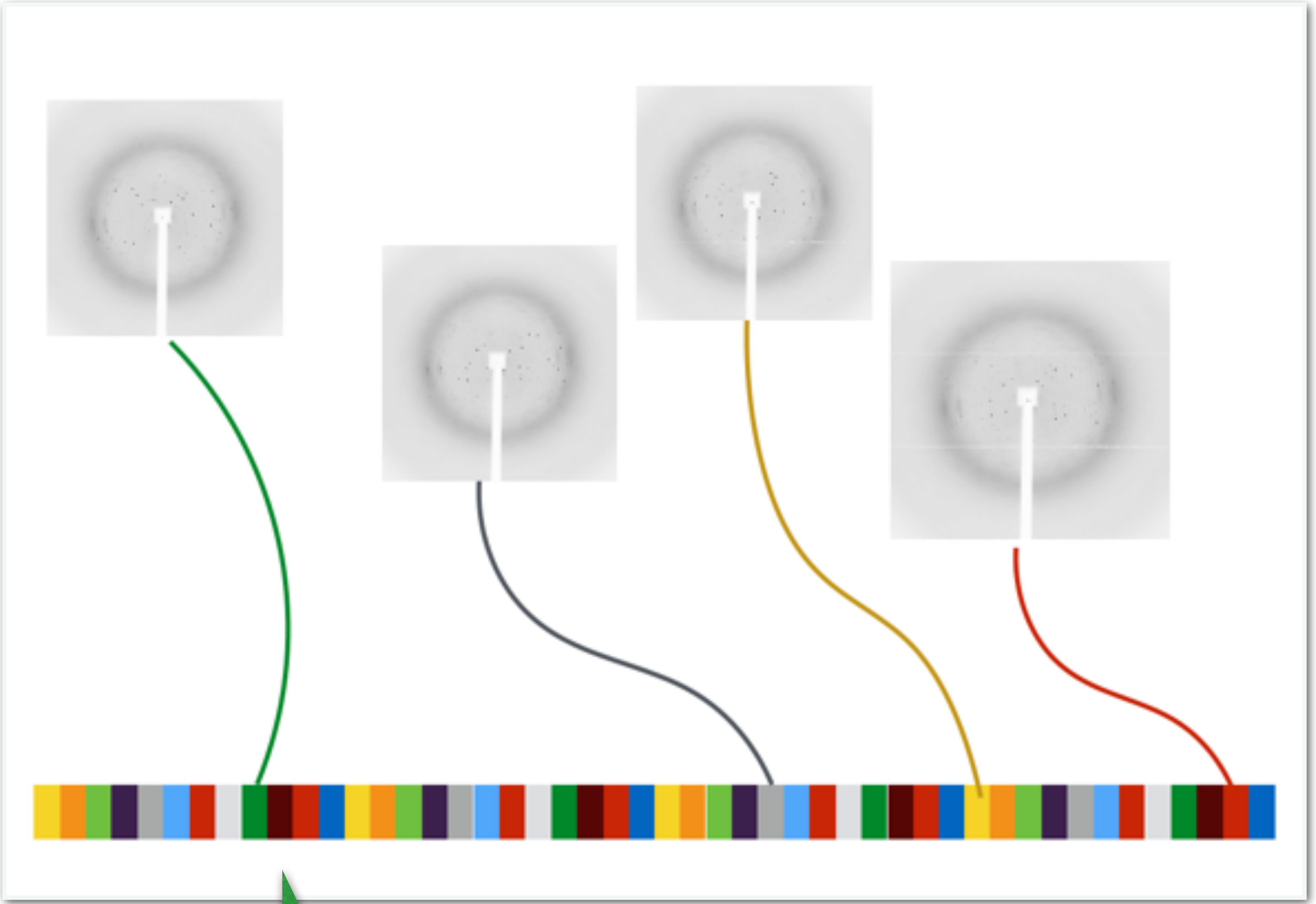
Nam et al., Cell, 2011

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

Four Essential Components of a biomedical data grid



Data Publication

Data Access

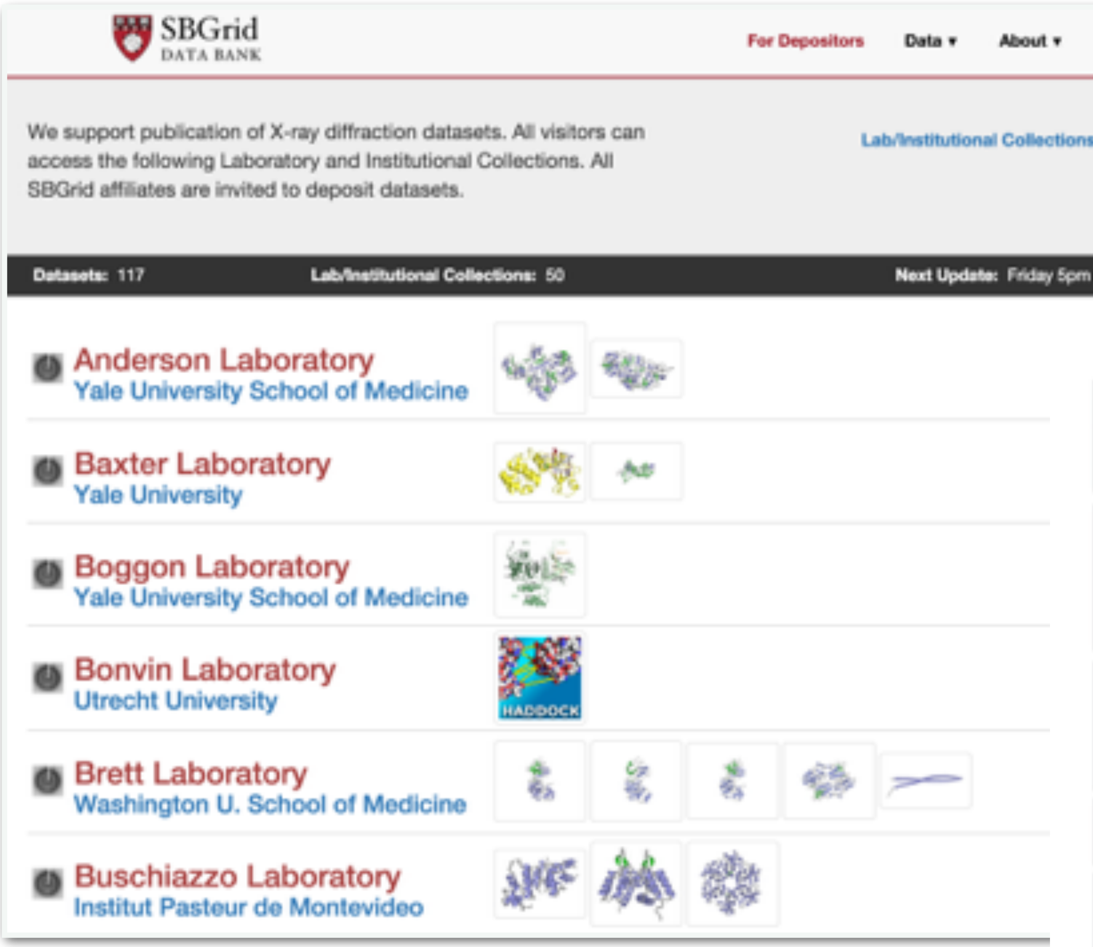
Live Analysis

Sustainability

Structural Biology Data Grid Website

data.sbgrid.org

Lab Collections



SBGrid DATA BANK

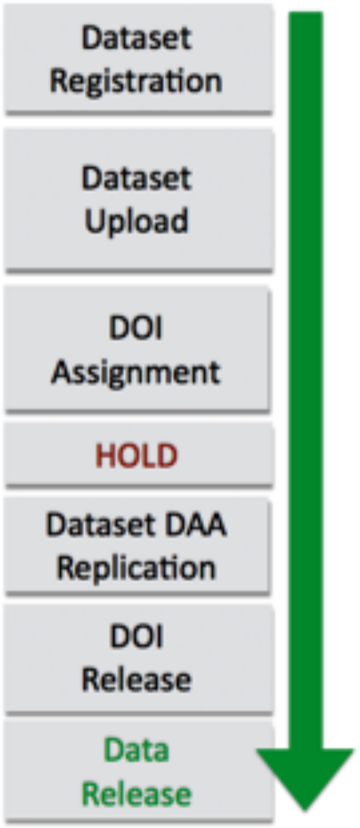
For Depositors | Data ▾ | About ▾ | Get Help ▾

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
- Buschiazzo Laboratory — Institut Pasteur de Montevideo



Dataset Registration

Dataset Upload

DOI Assignment

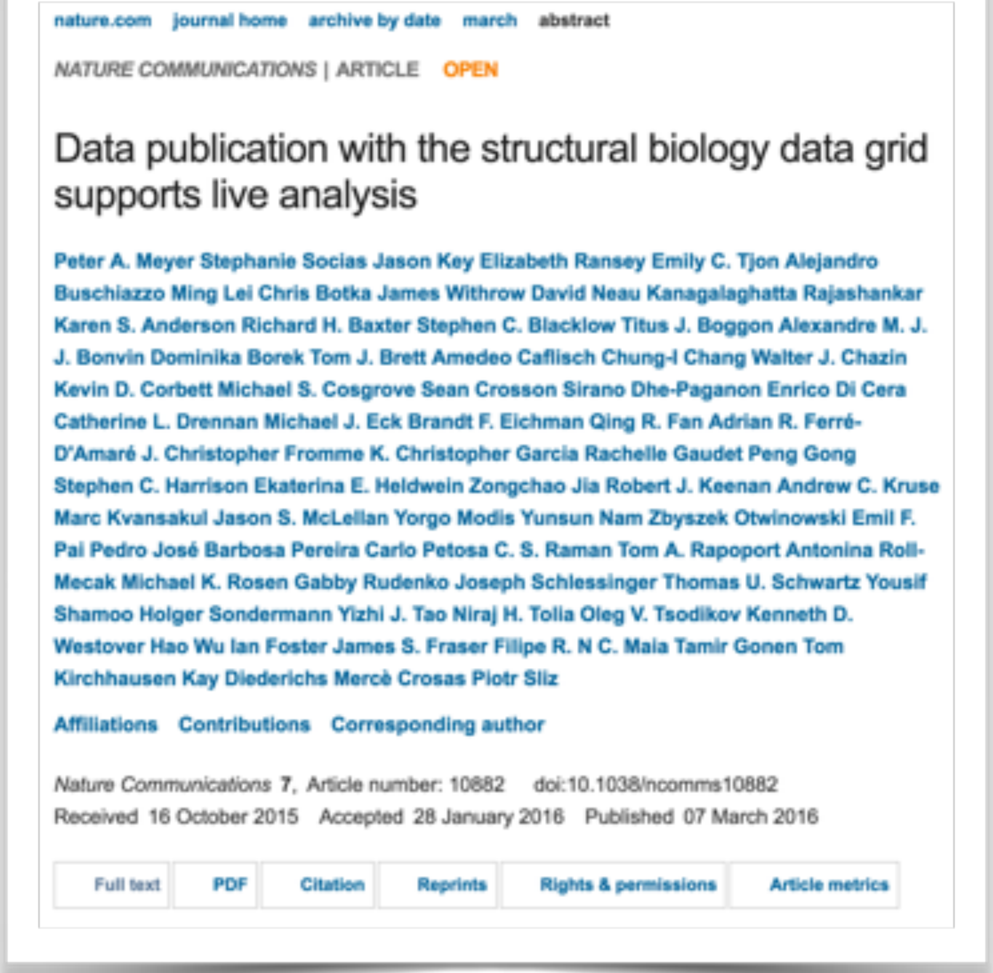
HOLD

Dataset DAA Replication

DOI Release

Data Release

doi:10.1038/ncomms10882



nature.com | journal home | archive by date | march | abstract

NATURE COMMUNICATIONS | ARTICLE OPEN

Data publication with the structural biology data grid supports live analysis

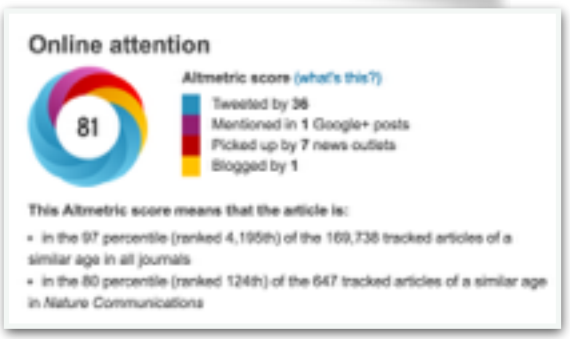
Peter A. Meyer Stephanie Socias Jason Key Elizabeth Ransey Emily C. Tjon Alejandro Buschiazzo Ming Lei Chris Botka James Withrow David Neau Kanagalaghatta Rajashankar Karen S. Anderson Richard H. Baxter Stephen C. Blacklow Titus J. Boggon Alexandre M. J. J. Bonvin Dominika Borek Tom J. Brett Amedeo Caflisch 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 J. Christopher Fromme K. Christopher Garcia Rachelle Gaudet Peng Gong Stephen C. 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 C. S. 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 James S. Fraser Filipe R. N C. Maia Tamir Gonen Tom Kirchhausen Kay Diederichs Mercè Crosas Piotr Sliz

Affiliations | Contributions | Corresponding author

Nature Communications **7**, Article number: 10882 doi:10.1038/ncomms10882

Received 16 October 2015 Accepted 28 January 2016 Published 07 March 2016

Full text | PDF | Citation | Reprints | Rights & permissions | Article metrics



Online attention

Altmetric score (what's this?)

81

- Tweeted by **36**
- Mentioned in **1** Google+ posts
- Picked up by **7** news outlets
- Blogged by **1**

This Altmetric score means that the article is:

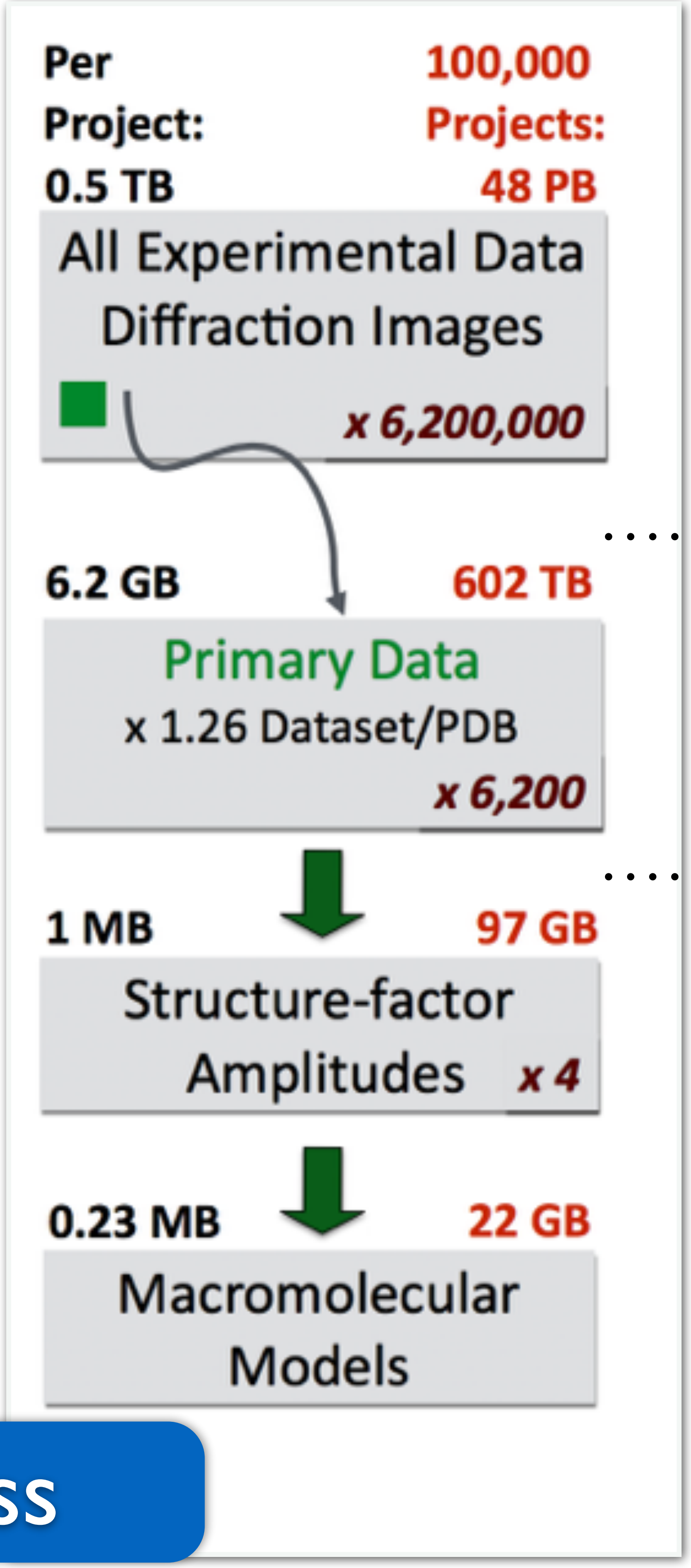
- in the 97 percentile (ranked 4,195th) of the 169,738 tracked articles of a similar age in all journals
- in the 80 percentile (ranked 124th) of the 647 tracked articles of a similar age in *Nature Communications*

Storage Requirements

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

SBDG:
110 datasets
~0.5TB

PDB:
100,000 models
0.3 TB



Per Project: 0.5 TB

100,000 Projects: 48 PB

All Experimental Data
Diffraction Images

x 6,200,000

6.2 GB

602 TB

Primary Data
x 1.26 Dataset/PDB

x 6,200

1 MB

97 GB

Structure-factor Amplitudes x 4

0.23 MB

22 GB

Macromolecular Models

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

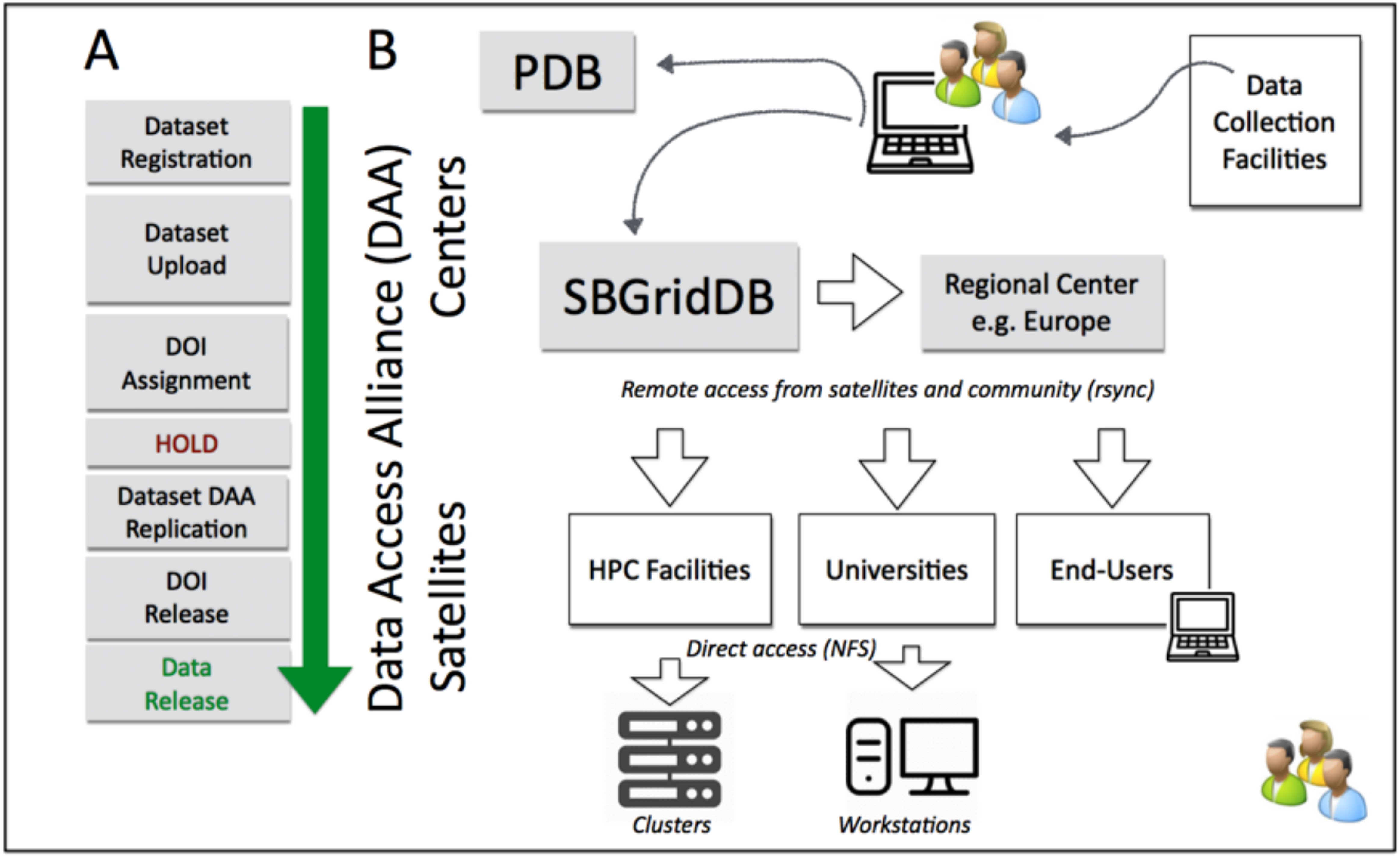
Molecular models and reduced datasets are stored in Protein Data Bank

WORLDWIDE PDB PROTEIN DATA BANK

Data Access

@SBGrid

Dataset Access Alliance:
Local access through a growing list of Satellites



National Data Service Pilot

@SBGrid

Dataset Access Alliance: Local access through a growing list of Satellites

UPPSALA
UNIVERSITET

Petrel Storage
at Argonne Labs

Orchestra Cluster
Harvard Medical School

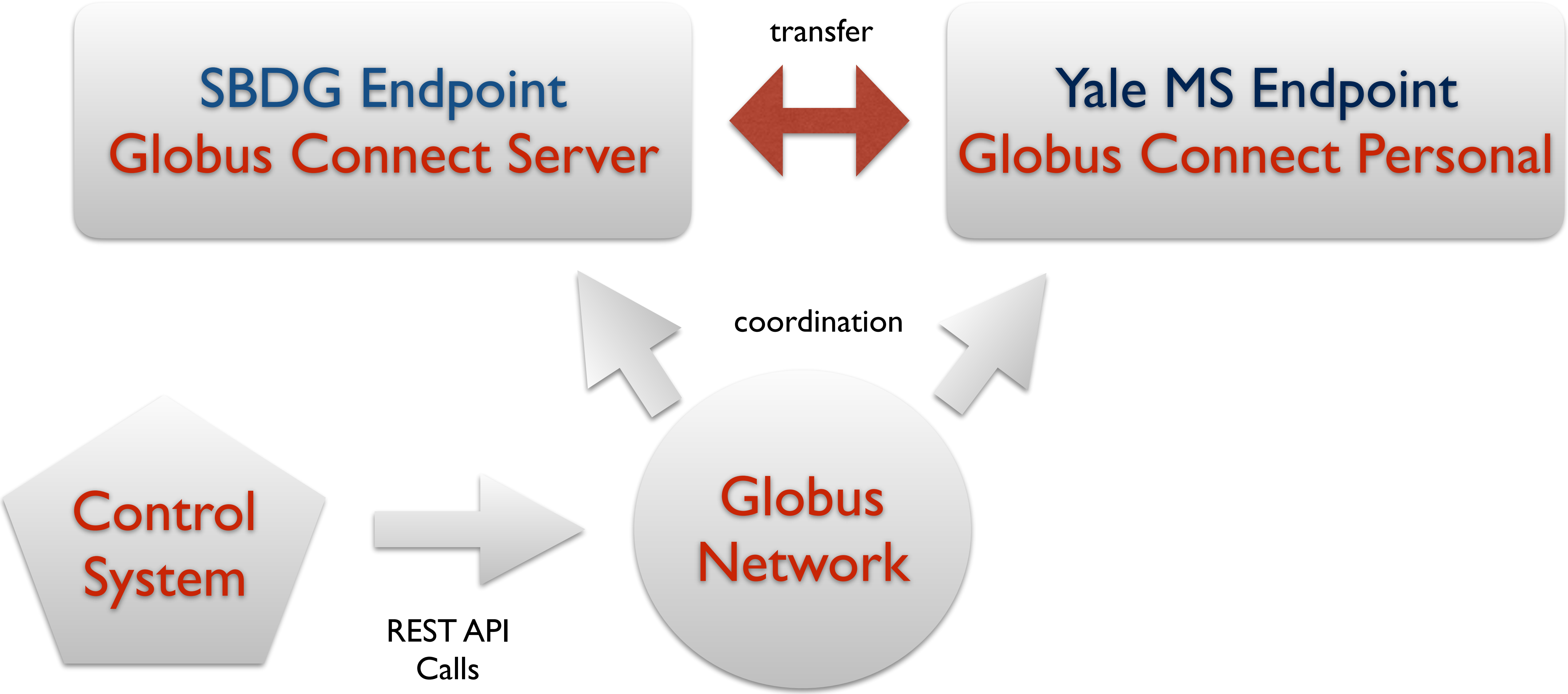
San Diego
Supercomputer
Center

INSTITUT PASTEUR
DE MONTEVIDEO

SIBS SHANGHAI INSTITUTES FOR BIOLOGICAL SCIENCES,
CHINESE ACADEMY OF SCIENCES

National Data Service Pilot

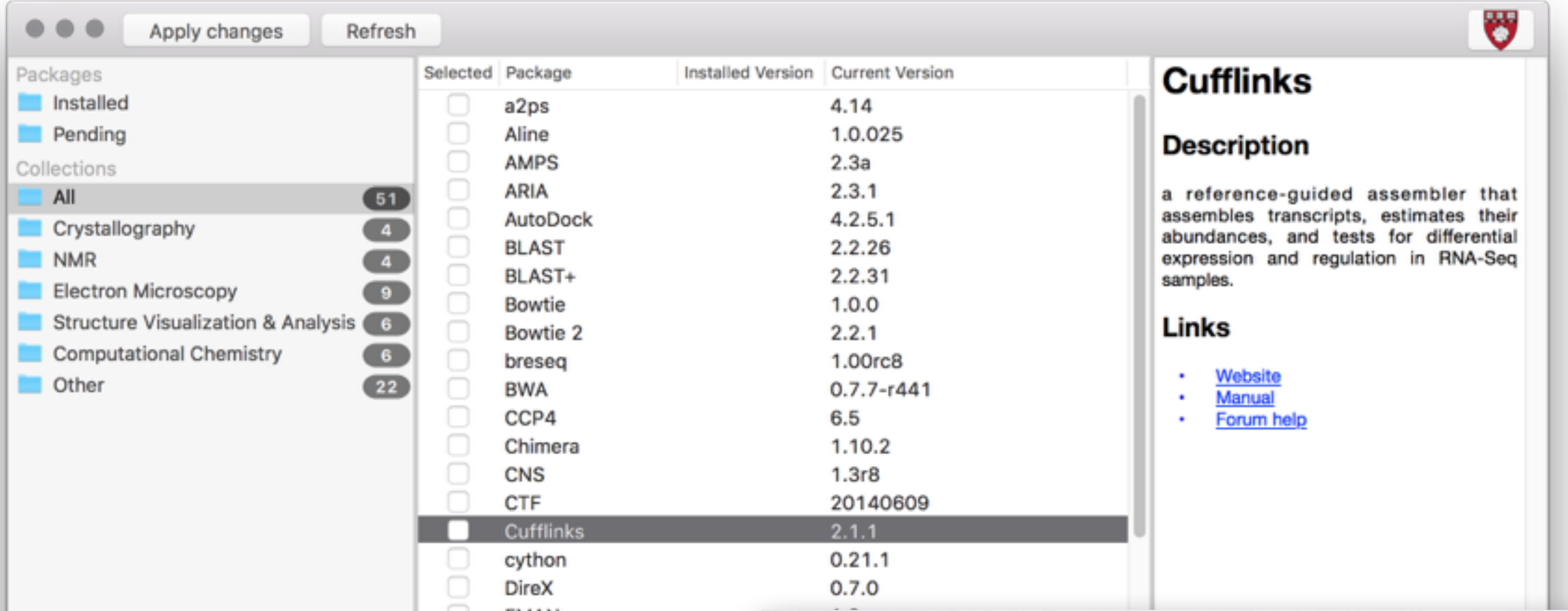
Dataset Access Alliance:
Local access through a growing list of Satellites



National Data Service Pilot

@SBGrid

Integration with Biomedical Software Support



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
☐	a2ps		4.14
☐	Aline		1.0.025
☐	AMPS		2.3a
☐	ARIA		2.3.1
☐	AutoDock		4.2.5.1
☐	BLAST		2.2.26
☐	BLAST+		2.2.31
☐	Bowtie		1.0.0
☐	Bowtie 2		2.2.1
☐	breseq		1.00rc8
☐	BWA		0.7.7-r441
☐	CCP4		6.5
☐	Chimera		1.10.2
☐	CNS		1.3r8
☐	CTF		20140609
☐	Cufflinks		2.1.1
☐	cython		0.21.1
☐	DireX		0.7.0
☐	EMAN		
☐	EMAN2		
☐	EPMR		
☐	FASTA		
☐	FREALIGN		
☐	GROMACS		
☐	HADDOCK		
☐	IGV		
☐	MAFFT		
☐	MGLTools		

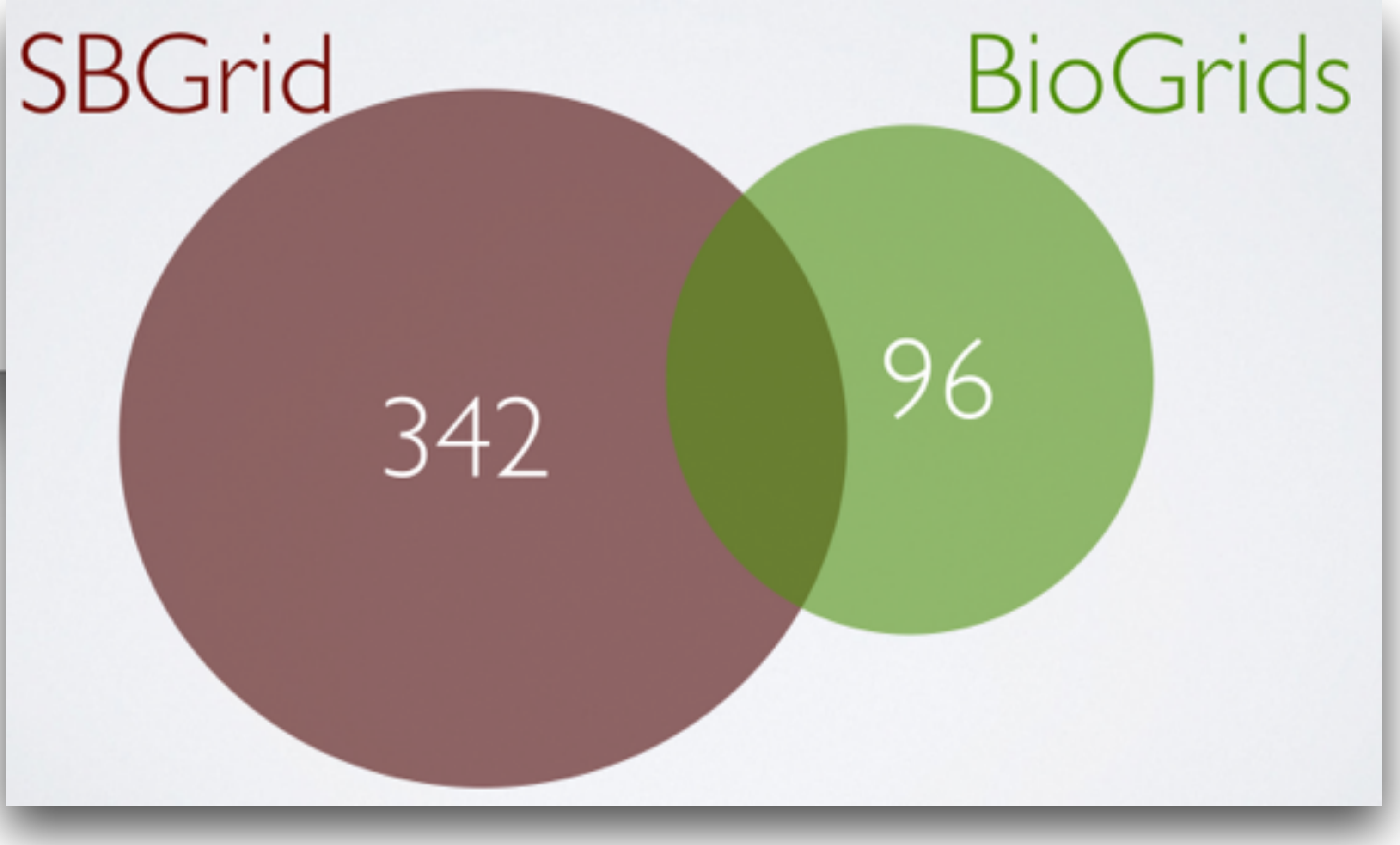
Cufflinks

Description

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

Links

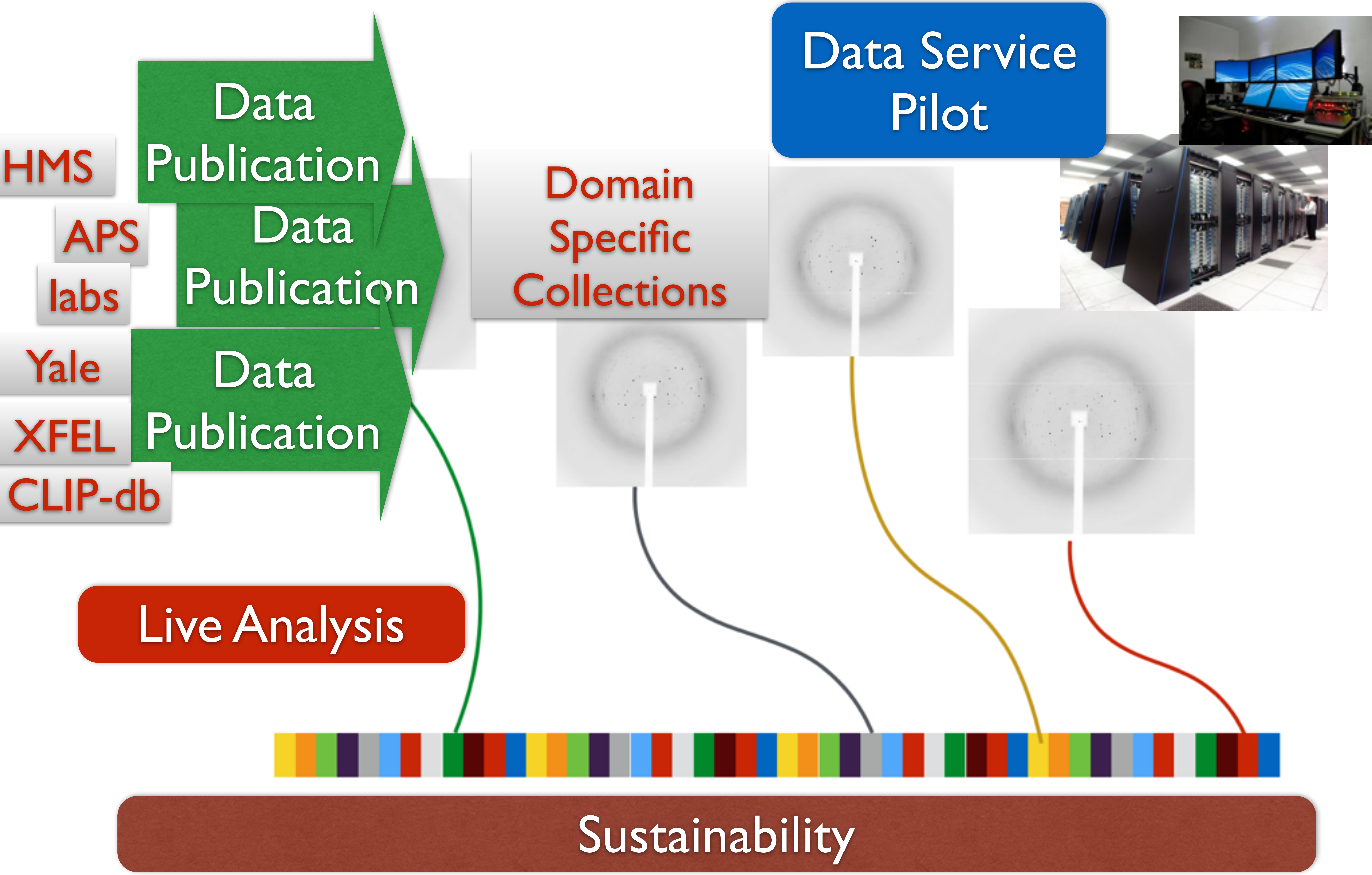
- Website
- Manual
- Forum help



sbgrid.org

biogrids.org

Summary



HMS

APS

labs

Yale

XFEL

CLIP-db

Data Publication

Data Publication

Data Publication

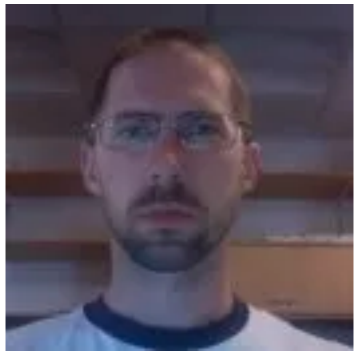
Domain Specific Collections

Data Service Pilot

Live Analysis

Sustainability

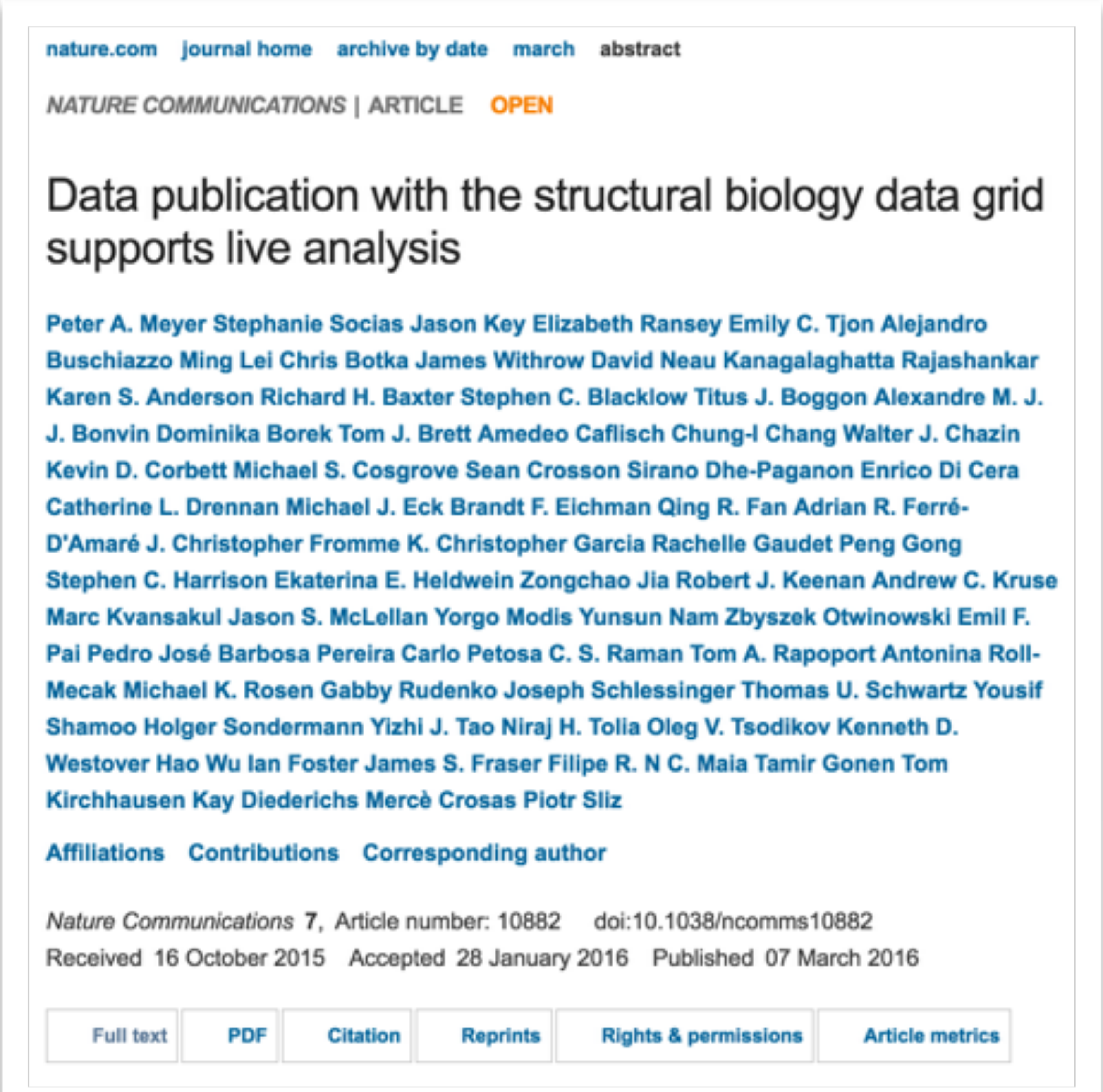
Stephanie Socias



Pete Meyer

Thank you

THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST



nature.com journal home archive by date march abstract

NATURE COMMUNICATIONS | ARTICLE OPEN

Data publication with the structural biology data grid supports live analysis

Peter A. 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 H. Baxter Stephen C. 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 J. Christopher Fromme K. Christopher Garcia Rachelle Gaudet Peng Gong Stephen C. 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 C. S. 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 James S. Fraser Filipe R. N C. Maia Tamir Gonen Tom Kirchhausen Kay Diederichs Mercè Crosas Piotr Sliz

Affiliations Contributions Corresponding author

Nature Communications 7, Article number: 10882 doi:10.1038/ncomms10882
Received 16 October 2015 Accepted 28 January 2016 Published 07 March 2016

- 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.

@SBGrid