THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST

Twitter: @SBGrid
YouTube: SBGridTV

Data Publication and Dissemination with the Structural Biology Data Grid

Piotr Sliz
Harvard Medical School

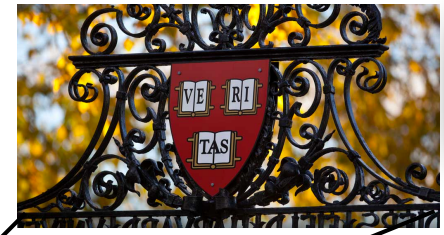
Merce Crosas
Harvard University

SBGrid Consortium
Harvard Medical School

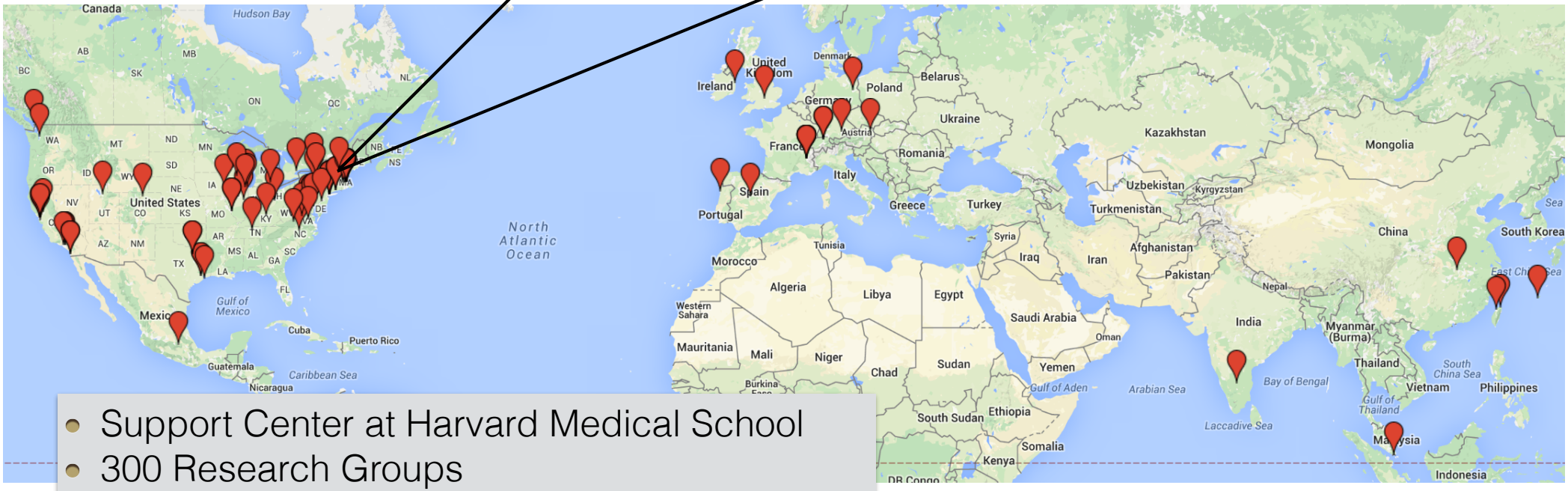
National Data Service
San Diego, CA, Oct. 20th, 2015

Background

SBGrid Consortium



Harvard Medical School



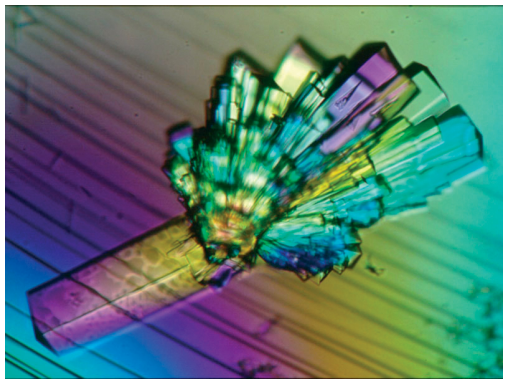
- Support Center at Harvard Medical School
- 300 Research Groups
- 13 Countries
- Long Term Sustainability: Membership Fee

Andrew Morin, Ben Eisenbraun, Jason Key, Paul C Sanschagrin, Michael A Timony, Michelle Ottaviano and Piotr Sliz Collaboration gets the most out of software. **eLife e01456: (2013).**

Andrew Morin, Jennifer Urban, Paul D Adams, Ian Foster, Andrej Sali, David Baker and Piotr Sliz. Shining Light into Black Boxes. **Science 6078: 159-160 (2012).**

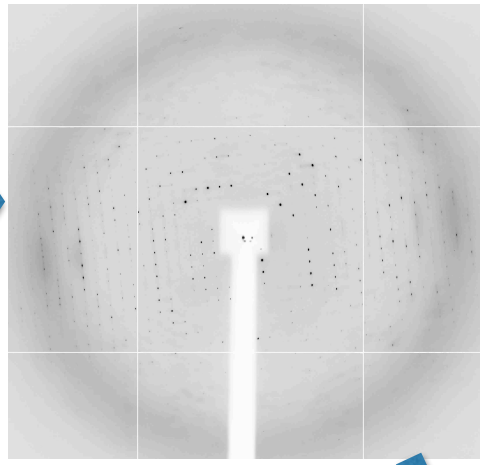
Experimental datasets in X-ray crystallography are acquired as Diffraction Images

Biological Samples
(ie. protein crystals)



Difficult to reproduce and handle

Experimental Data
(ie. diffraction images)

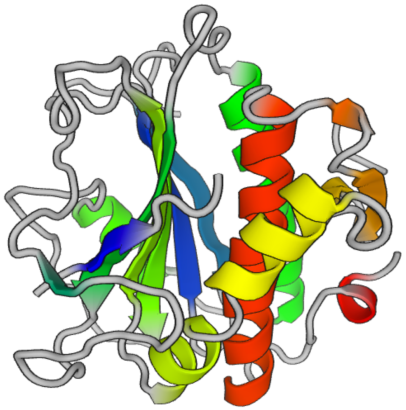


Collected at National Facilities

Primary Data

Secondary Data

Model
(ie. PDB files)



Built and refined by biologists.

- Primary Data used to derive structures are typically preserved on tapes/DVDs by researchers (~100,000 datasets lost)
- Non-primary data is often left behind at synchrotrons, or stored locally on poorly annotated media

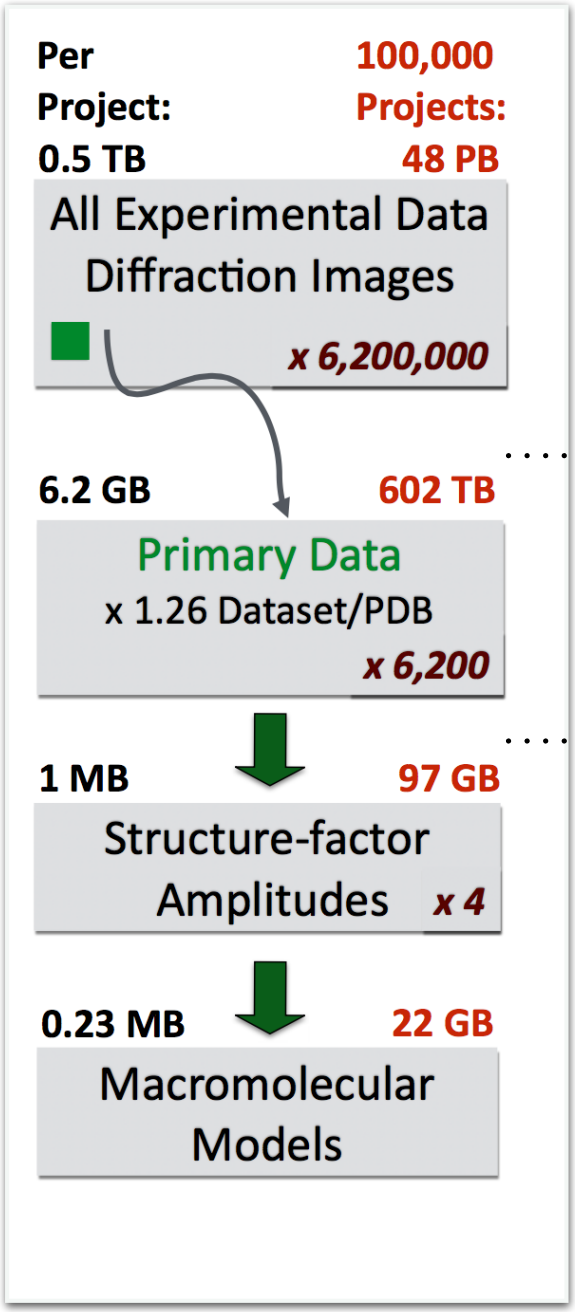
@SBGrid

Community 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



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

@SBGrid

Benefits of access to experimental datasets. Paradigm shift: Continuously Improving Models

- Datasets can be reprocessed and used to derive structures with new software
- Improved criteria can be applied (e.g. resolution limits such CC1/2, Karplus and Diederichs, 2012, or anisotropic correction)
- Datasets can be used to validate questionable structures (e.g. carboplatin, ABC transporter MsbA, complement pathway C3b)
- Anisotropic diffuse scattering signals can be analyzed to provide information about dynamics (e.g. Cyclophilin A)
- Education, teaching and software development.

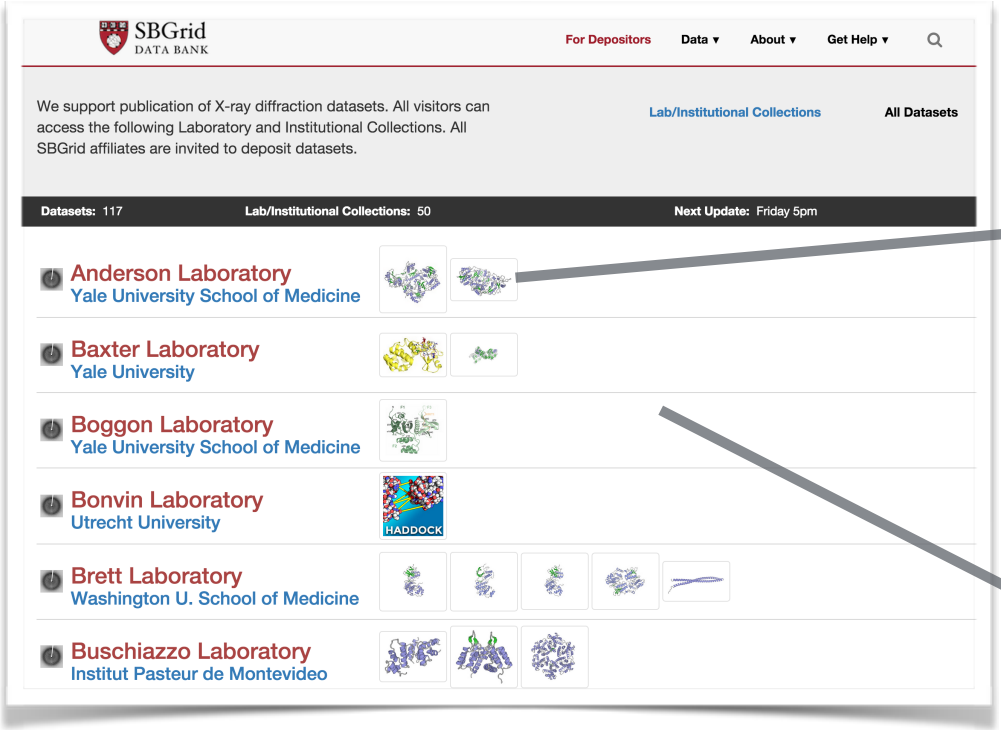
Community Calls for preservation of XC Datasets:

- Stokes-Rees, I., Levesque, I., Murphy, F.V., Yang, W., Deacon, A., and Sliz, P. (2012). Adapting federated cyberinfrastructure for shared data collection facilities in structural biology. J Synchrotron Radiat 19, 462–467.
- Terwilliger, T.C., and Bricogne, G. (2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. Acta Crystallogr. D Biol. Crystallogr. 70, 2533–2543.
- Terwilliger, T.C. (2014). Archiving raw crystallographic data. Acta Crystallogr D Biol Crystallogr.
- Guss, J.M., and McMahon (2014). How to make deposition of images a reality. Acta Crystallogr. D Biol. Crystallogr. 70, 2520–2532

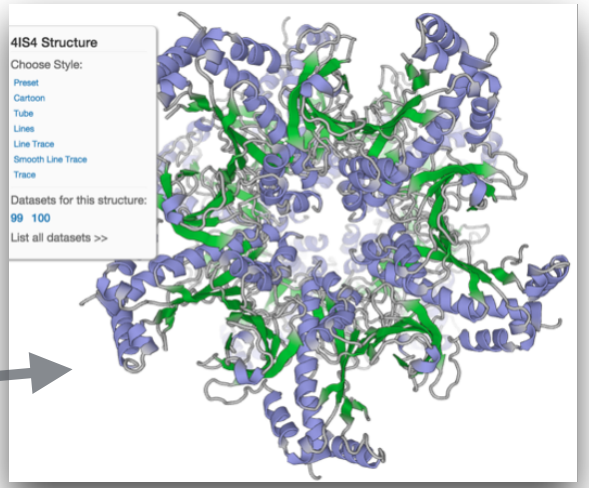
Implementation

Structural Biology Data Grid Portal.
URL: data.sbgrid.org

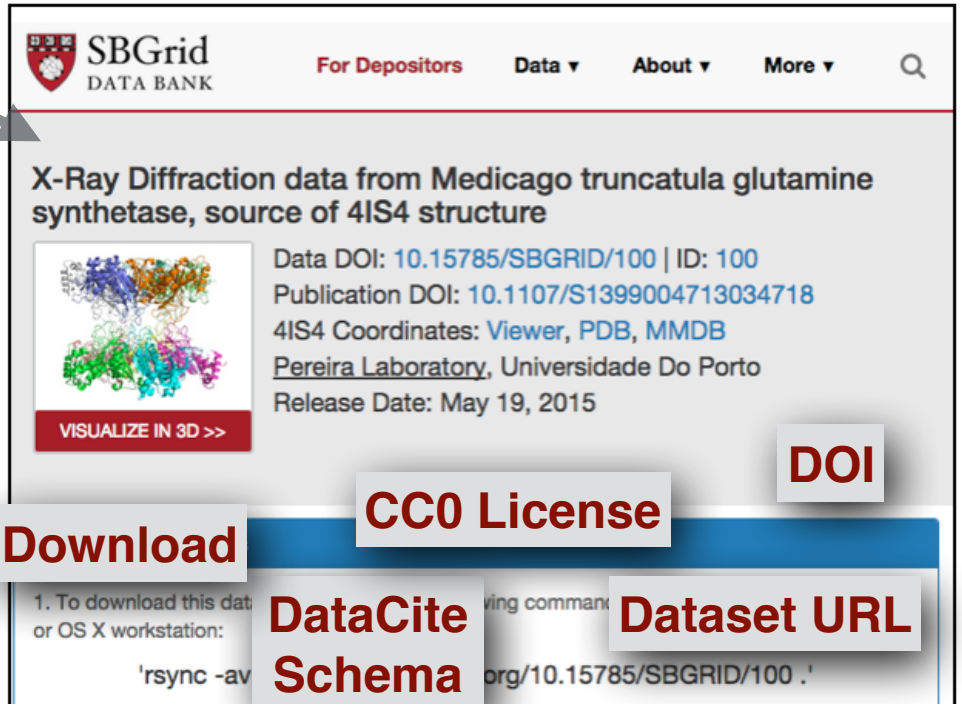
Laboratory Collections



Viewer



Persistent Dataset Pages



- Bootstrap front-end framework along with custom CSS, HTML, and JavaScript
- Python, utilizing Django 1.7
- PostgreSQL 9.3 for metadata storage

Data Access Alliance: Remote Data Access
Download from DAA Centers

US (Harvard Medical School)

```
rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/164 .
```

South America (Institut Pasteur de Montevideo, Uruguay)

```
rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/164 .
```

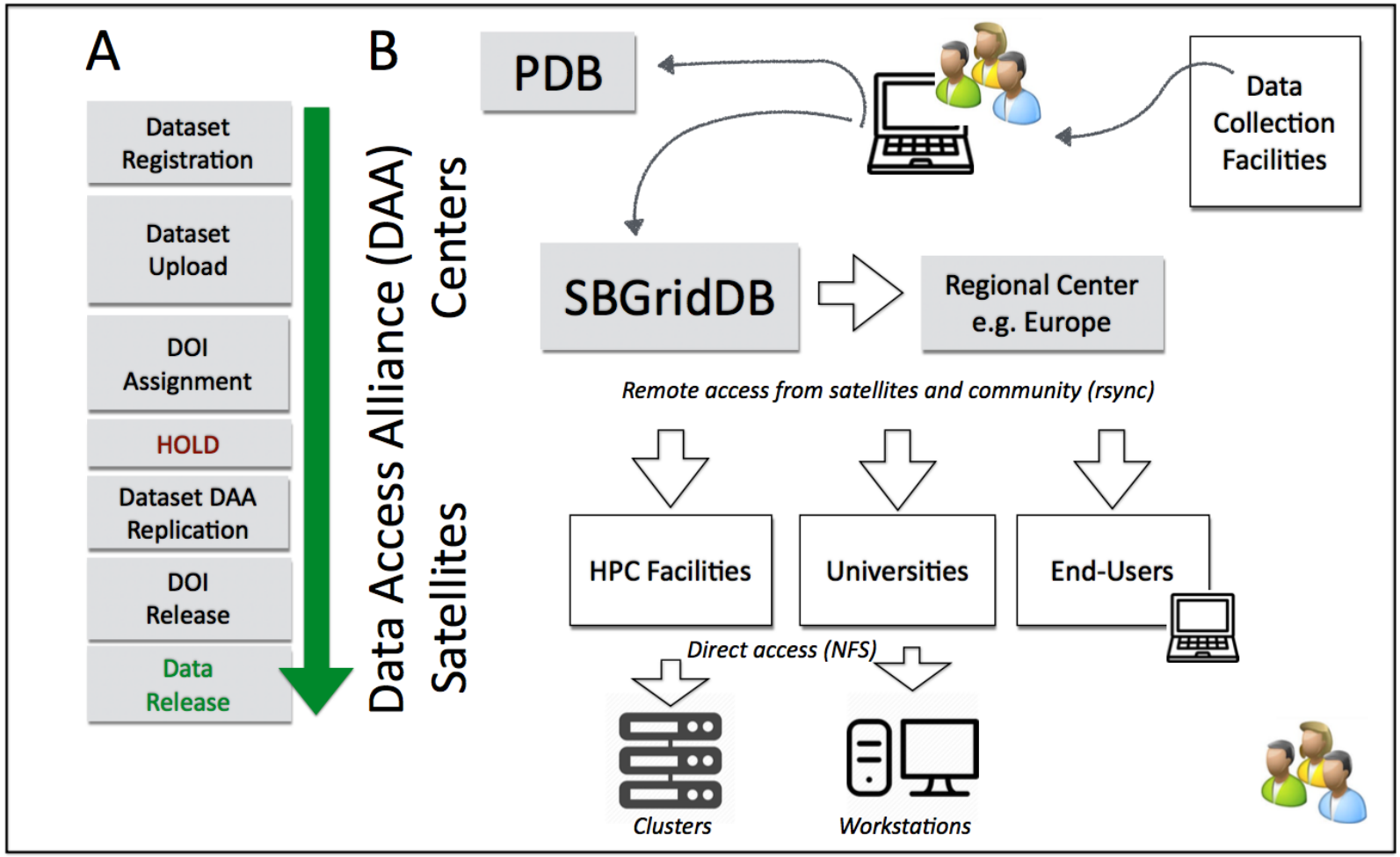
Europe (Uppsala University, Sweden)

```
rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/164 .
```

Asia (Shanghai Institutes for Biological Sciences, China)

```
rsync -av rsync://TBD/10.15785/SBGRID/164 .
```

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

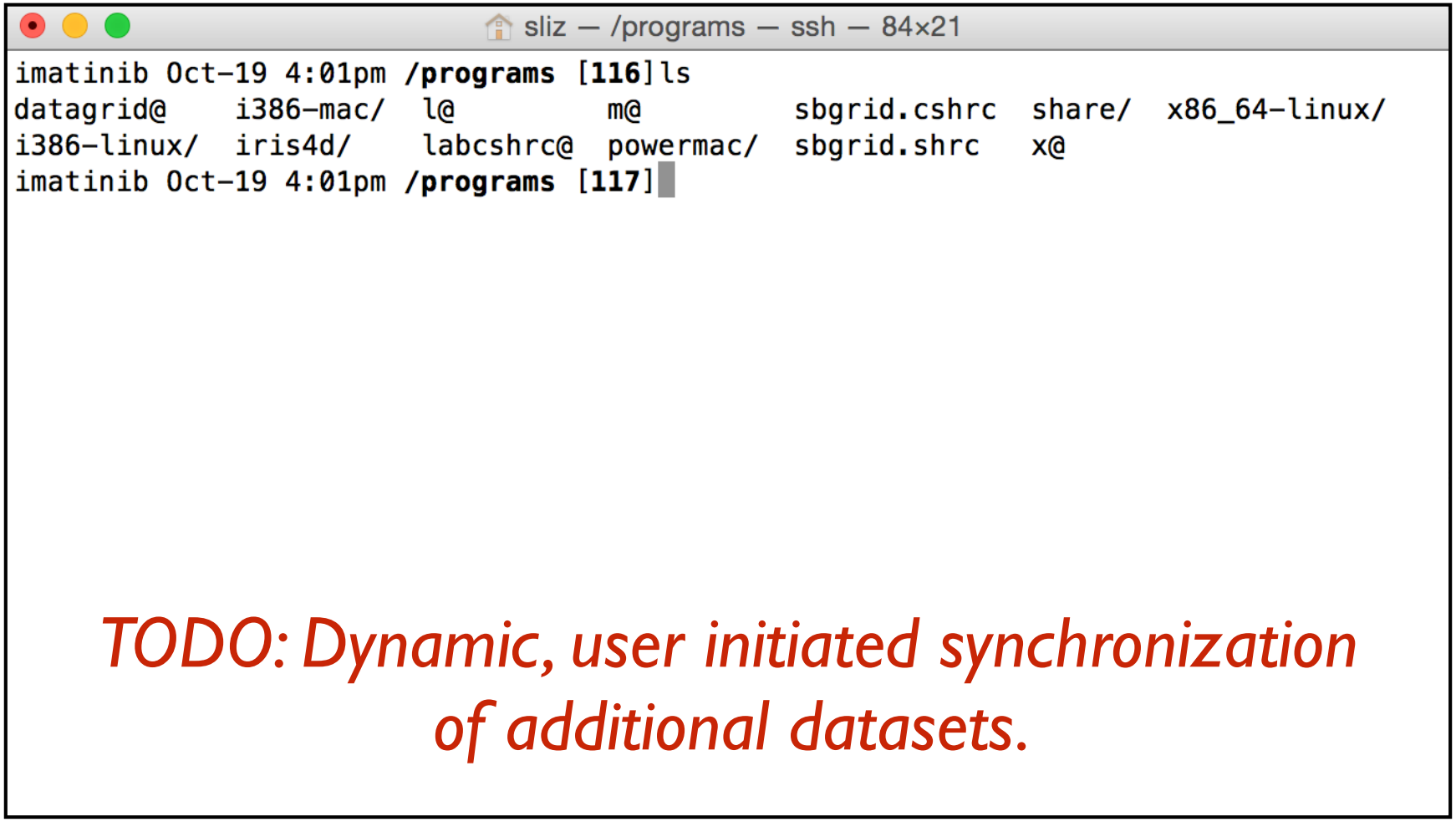


TODO: Access to the National Compute Infrastructure

@SBGrid

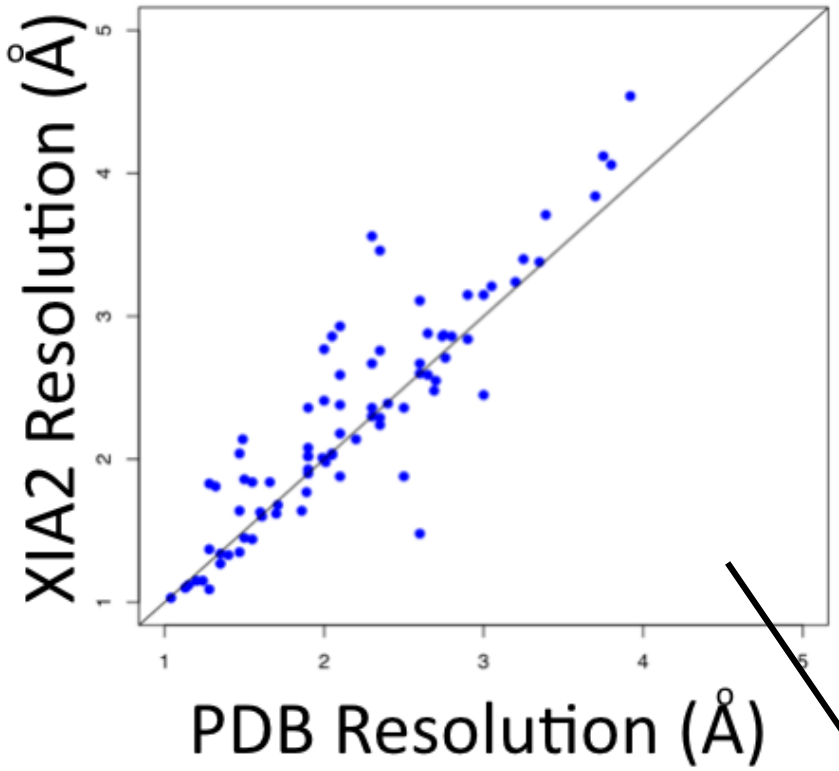
Local Access:
Data and Software Integration

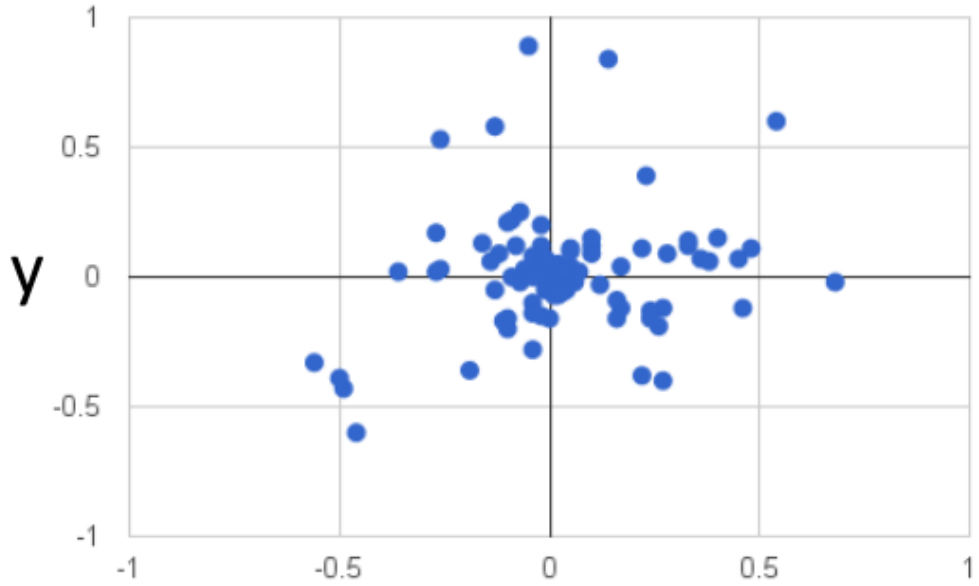
Example: Yale Medical School

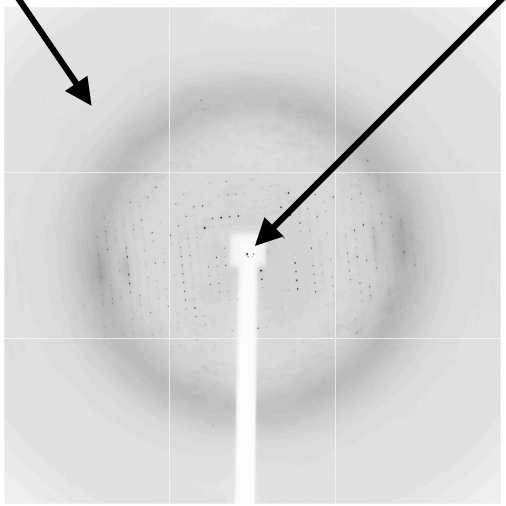


TODO: Dynamic, user initiated synchronization of additional datasets.

Post-deposition review (internal and external)
XIA2 and DIALS

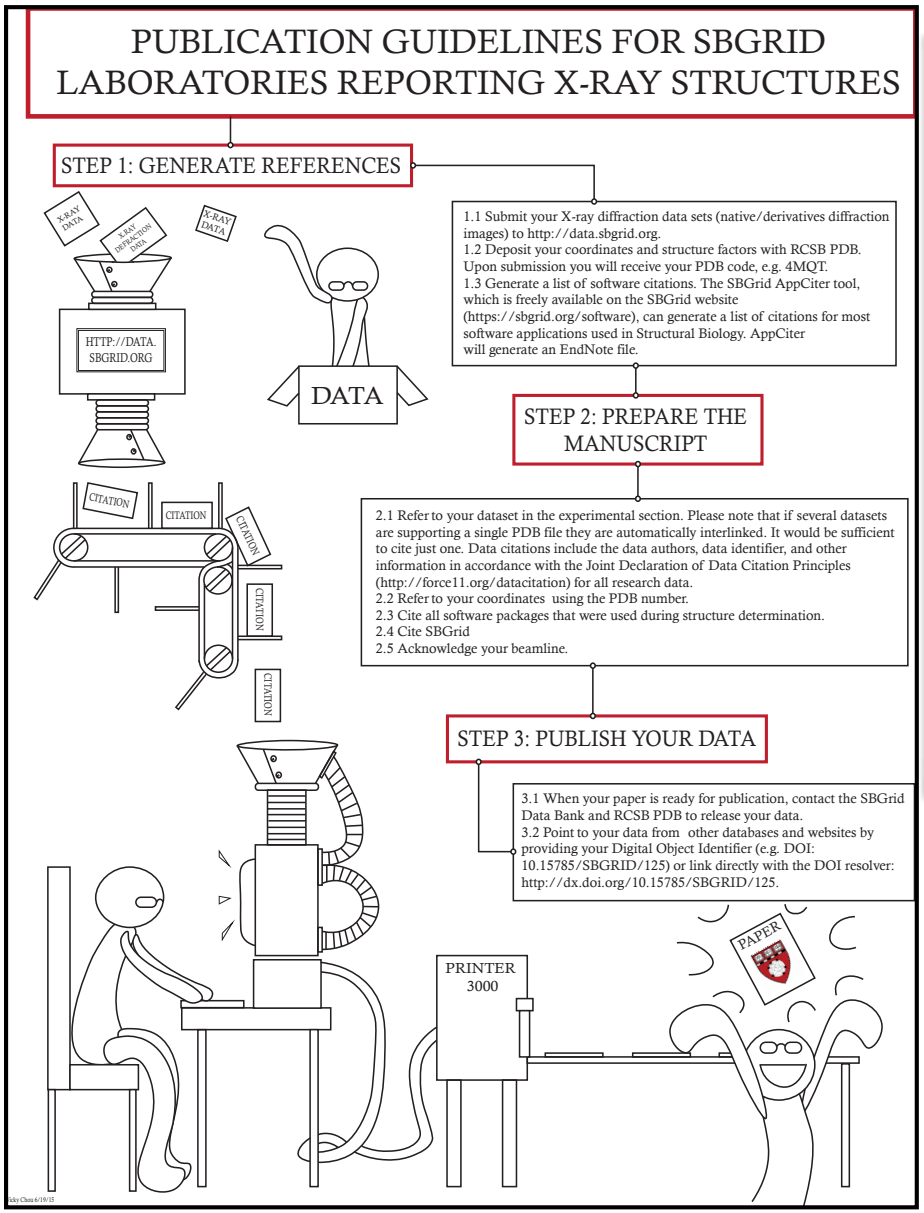


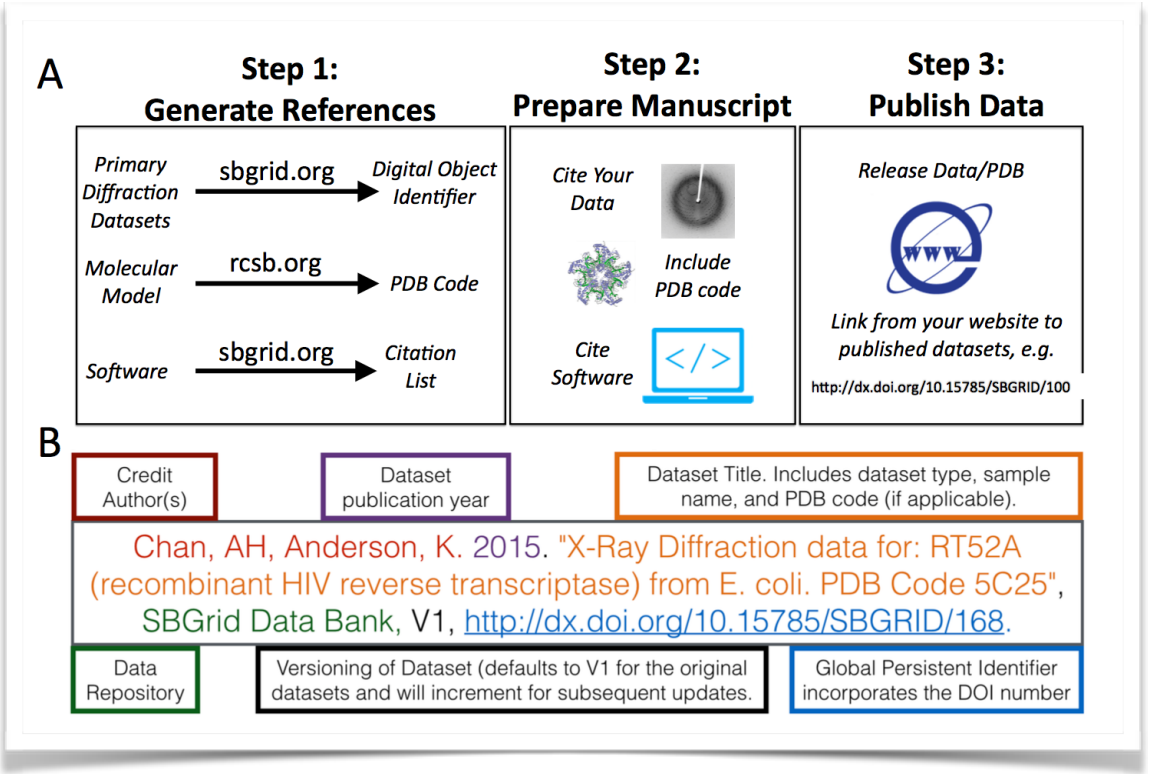




ONGOING:
External Validation
at APS.

Comprehensive Publication Workflow: Citations for Data + Models + Software

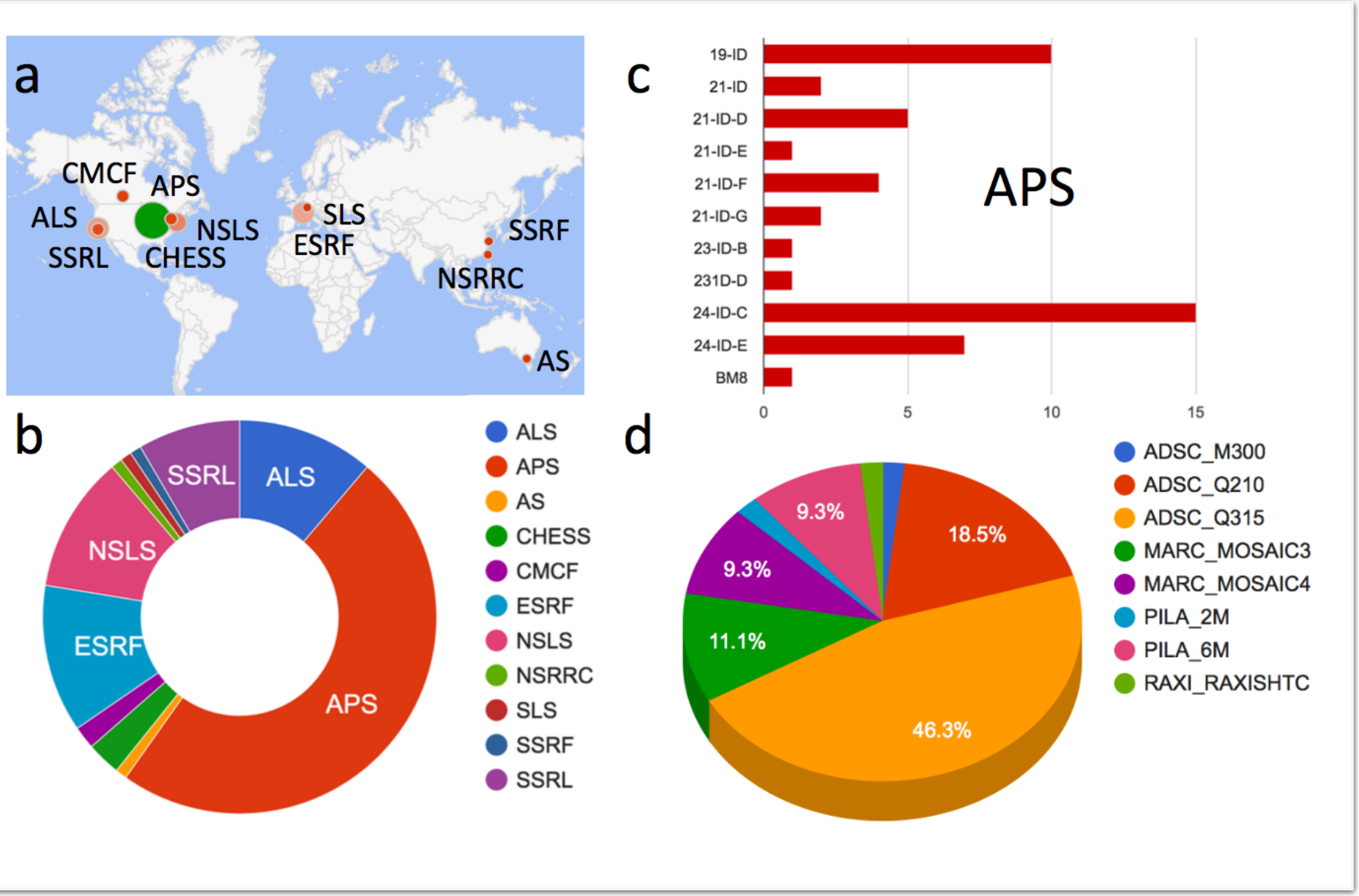




Data Citation Synthesis Group:
Joint Declaration of Data Citation Principles

@SBGrid

Data collection statistics for the pilot subset of 111 X-ray diffraction datasets.



@SBGrid

Structural Biology Data Grid as community "Dropbox" support for additional "huge" data types

- MicroED (electron diffraction from micro crystals). Deposited datasets originate from Tamir Gonen laboratory at HHMI Janelia Farm. (5GB/dataset)
- Molecular Dynamics. 2 μs simulation to demonstrate OGT peptide loading was completed on an MPI cluster with Desmond software.
- Decoy Datasets for NMR computations. (50GB/dataset)
- XFEL Datasets. SBDB will replicate datasets from the Coherent X-ray Imaging Data Bank images. (~100GB/dataset)
- Lattice Light-Sheet Microscopy (new instruments at Harvard Medical School) - (

TODO: Community curators for new types of datasets + API to support external validation and annotation

Summary

- We have established a diverse collection of ~120 X-ray Diffraction Images (support 70 journal publications, created by 50+ research groups, CC0)
- FORCE11 "compliant" - DOIs, DataCite Schema and data citations recommendations.
- **Data Preservation/Access** through Data Access Alliance: SBGrid Consortium will utilize storage resources of its members. Make the data immediately accessible for computations.
- **Data Review:** Dynamic post-deposition review, completed at Harvard and by other members of our community.
- SBDG can support **other types of biomedical data** (dynamic review process and data descriptors will need to be developed)
- The data deposition system complements SBGrid's mission to maintain a community-wide research-software infrastructure (e.g. AppCiter for software citation, software collection, collection of ~60 software tutorials).

Discussion

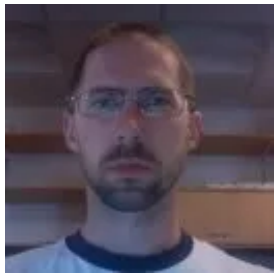
Project Specific:

- **Technology:** Transition to a Research Data Management System (combination of Dataverse and Globus Toolkit). Handling of user interfaces, Data Access Alliance systems, data validation, ORCID.

Potential Topics to be addressed by NDS Pilot:

- **Policy:** Advocacy for community wide data deposition (Nature, Cell, ACA, NIH, NSF, DoE, Synchrotrons). Where to start? Workshop/Editorial articles?
- **User education**: FORCE11, webinar or animation. This would not have to be data type specific.
- **Data Replication:** Integration with US Cyberinfrastructure/NDS sites for compute access and preservation sites. Uniform technology to replicate data.
- Workforce and User Groups. Workforce training.
- **Funding Models:** Long-term financial support for development and operations (e.g. Consortium, Institutions). Pressures to establish a self-sustainable system.

Stephanie Socias



Pete Meyer

Thank you

Mercè Crosas

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**

- 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