The Brain Image Library

Derek Simmel, dsimmel@psc.edu

iRODS User Group Meeting 2018
Durham, North Carolina, U.S.A.
June 7, 2018

The Brain Image Library

- Introduction
    - Pittsburgh Supercomputing Center (PSC)
    - NIH BRAIN Initiative
    - BRAIN Initiative Cell Census Network (BICCN)
- Brain Image Library (BIL)
    - Team
    - Motivation, Vision, and Goals
    - Communities Served
    - Technologies & Data Management with iRODS
    - Challenges

Carnegie Mellon University

PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh

Pittsburgh Supercomputing Center

Joint effort of Carnegie Mellon University and the University of Pittsburgh

32 years of national leadership in:
- High-performance computing and data analytics (HPC, HPDA)
- 19 high-performance computers, including 9 that were/are "serial #1"
- Research groups: Artificial Intelligence & Big Data, Biomedical Applications, Public Health Applications, User Support for Scientific Applications, Networking, Security
- Software architecture, implementation, and optimization
- Networking and network optimization
- Enabling ground-breaking science, computer science, and engineering
- Leading research in AI, biology, public health, neuroscience, filesystems, networking, HPC software engineering, chemistry, materials science, engineering, physics, statistics, …

Supported by: NSF, NIH, DOE, DoD, the Commonwealth of Pennsylvania, D. E. Shaw Research, National Energy Technology Laboratory, Bill and Melinda Gates Foundation, UNICEF, USAID, Grable Foundation, Buhl Foundation, and industry

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh

PSC's Multiple Roles

National service provider for research and discovery
- Bridges, Anton 2, Brain Image Library, Compass, XSEDE,…

Research institution centered on computational and data science
- 25+ currently active projects

Education and training
- Courses, national & local workshops

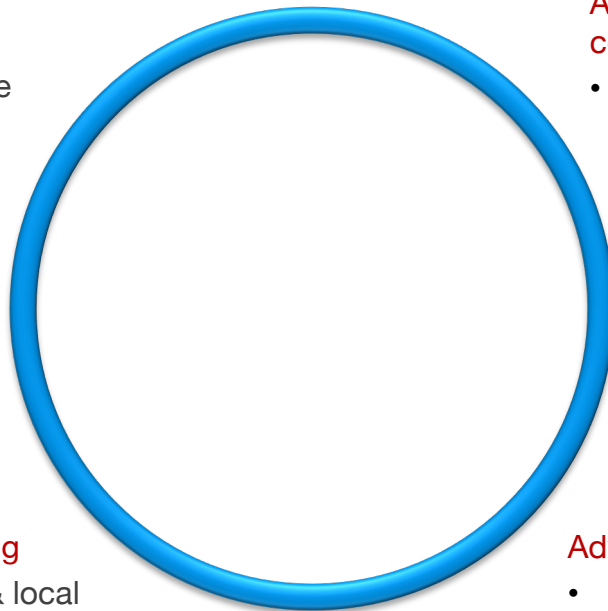**Active member in the CMU and Pitt communities**
- Research collaborations, interns

Networking
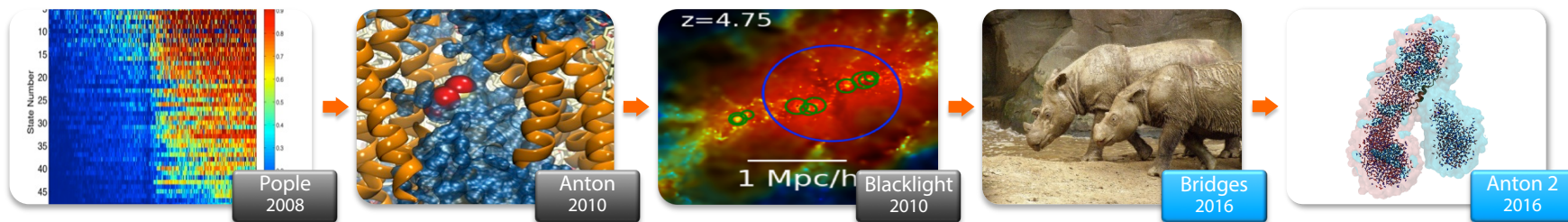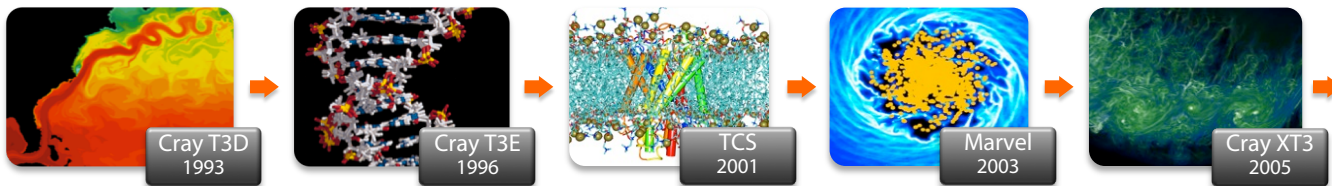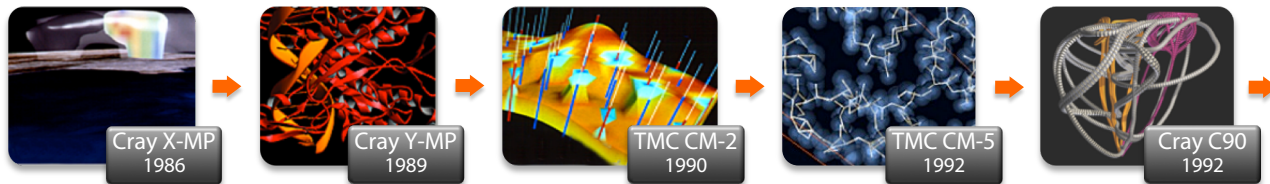- Service provider (3ROX), research

Advise and support industry
- Training, access to advanced resources, collaborative research

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh

PSC's Systems Enable New Science

Cray X-MP 1986 → Cray Y-MP 1989 → TMC CM-2 1990 → TMC CM-5 1992 → Cray C90 1992 → Cray T3D 1993 → Cray T3E 1996 → TCS 2001 → Marvel 2003 → Cray XT3 2005 → Pople 2008 → Anton 2010 → Blacklight 2010 → Bridges 2016 → Anton 2 2016

Carnegie Mellon University

PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh

NIH BRAIN Initiative

- U.S. National Institutes of Health (NIH) *Brain Research through Advancing Innovative Neurotechnologies* (BRAIN) Initiative
  - https://www.braininitiative.nih.gov
  - Funding to support infrastructure development and investigation in:
    - Cell Type
    - Circuit Diagrams
    - Monitoring Neural Activity
    - Interventional Tools
    - Theory and Data Analysis Tools
    - Human Neuroscience
    - Integrated Approaches

Carnegie Mellon University

PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh

NIH BRAIN Initiative Cell Census Network (BICCN)

2017: $250M awarded to 11 projects

- **Brain Cell Data Center**
  - A Community Resource for Single Cell Data on the Brain

- **Mouse Brain Cell Census Center**
  - A Comprehensive Center for Mouse Brain Cell Atlas
  - A Comprehensive Whole-Brain Atlas of Cell Types in the Mouse
  - Center for Epigenomics of the Mouse Brain Atlas

- **Mouse Brain Cell Census Collaboratory**
  - Collaboratory for atlasing cell type anatomy in the female and male mouse brain
  - Anatomical characterization of neuronal cell types of the mouse brain

- **Human and Nonhuman Primate Brain Cell Census Collaboratory**
  - A cellular resolution census of the developing human brain
  - A multimodal atlas of human brain cell types
  - A molecular and cellular atlas of the marmoset brain

- **BRAIN Initiative data archives for access to raw data:**
  - A confocal fluorescence microscopy brain data archive   <-- This is us: **The Brain Image Library**
  - A BRAIN Initiative resource: The neuroscience multi-omic data archive

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh

The Brain Image Library

- The Brain Image Library (BIL) is one of two NIH R24 awardees charged with collecting, archiving, and making available raw data and associated metadata from participating U.S. brain science laboratories

- The Brain Image Library
  - Collects and stores raw image data and metadata (10+PB)
    - Anatomy & morphology image data
    - Spatial Transcriptomics data
  - Enables computation and remote visualization using BIL data
  - Enables queries and access across BICCN projects for image data

- BIL is housed at the Pittsburgh Supercomputing Center (PSC)
  - PSC provides extensive knowledge, experience and services in systems deployment, operations, networking, computational methods and tools, and user support
  - PSC computational resources will be able to access BIL data directly

Carnegie Mellon University

PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh

BIL Team: Pittsburgh Supercomputing Center

Alex Ropelewski

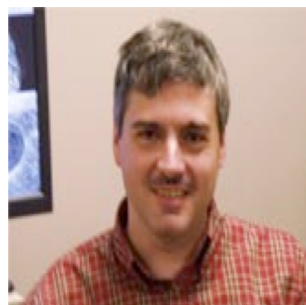Jacob Czech

Art Wetzel

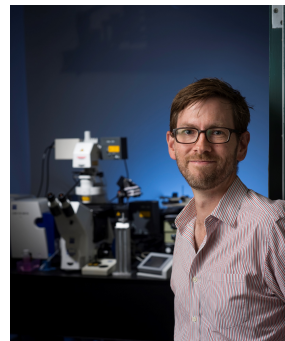Sean Deitrich

Derek Simmel

Kathy Benninger

Greg Hood

Shandra Williams

BIL Team: Carnegie Mellon University

- Molecular Biosensor and Imaging Center
  - Expertise in fluorescent probe design and validation and quantitative microscopy
  - Development of new tools for establishing cell-type specific connectivity in the rodent nervous system.


Marcel Bruchez

Greg Fisher

BIL Team: University of Pittsburgh

- University of Pittsburgh Center for Biologic Imaging
  - One of the largest and best equipped microscope based imaging centers in the country
  - New instrumentation developments in very fast imaging methods
  - Imaging high-resolution, large-scale volumetric data sets derived from static, cleared samples


Simon Watkins

Alan Watson

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh
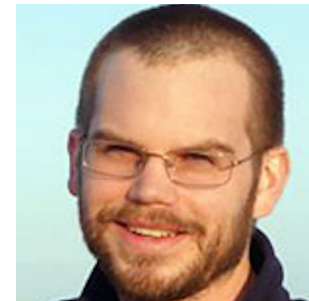
- Advancements in the field of microscopy have pushed the boundaries of what was once possible:
  - Goal is now to study systems at all levels of resolution from the single molecule to the whole animal at great depths of resolution
- To fully understand the brain, we need to:
  - Build, annotate and navigate 3D image volumes constructed from serial optical slices
  - Capture and correlate data from the full depth of the tissue

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh
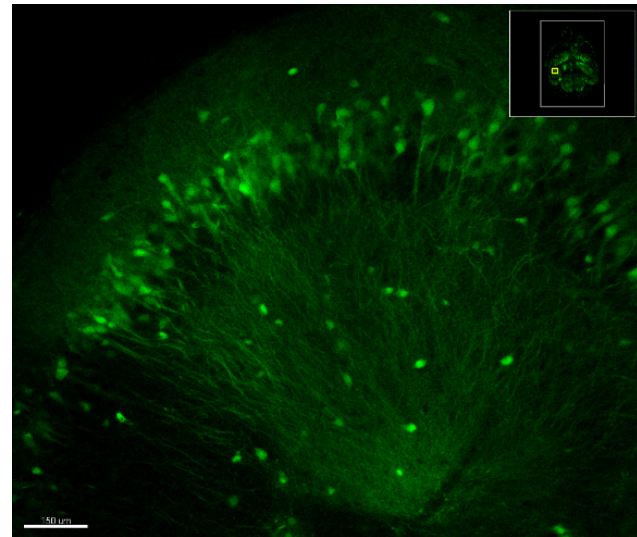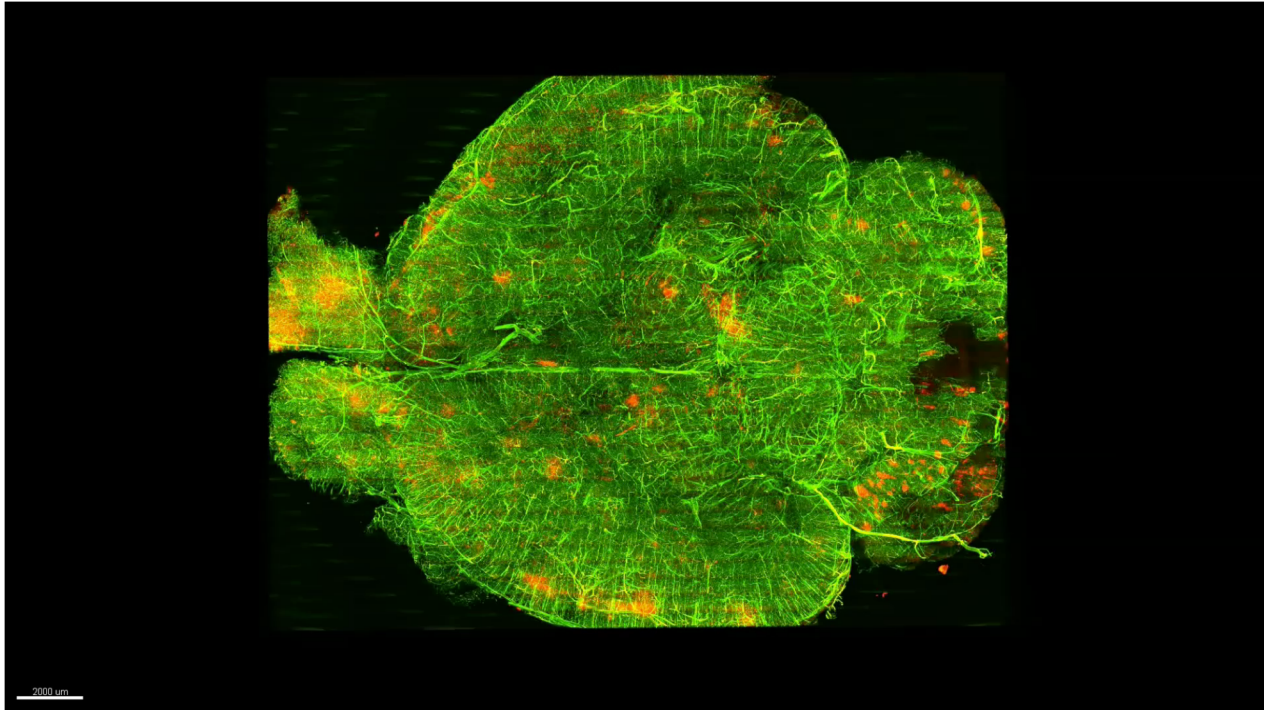
Big Data Problem

- New image capture systems are extremely fast

    - 4 terabytes/hour

    - Ribbon is 1000 pixels wide

    - Total image size is 70,000x100,000 pixels

- Mouse brain: 10TB

- Marmoset brain: ~800TB

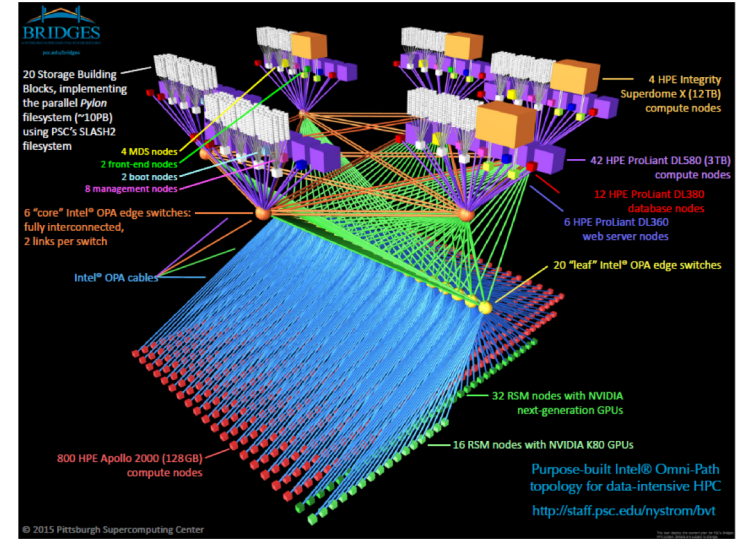- Human brain: ~1EB
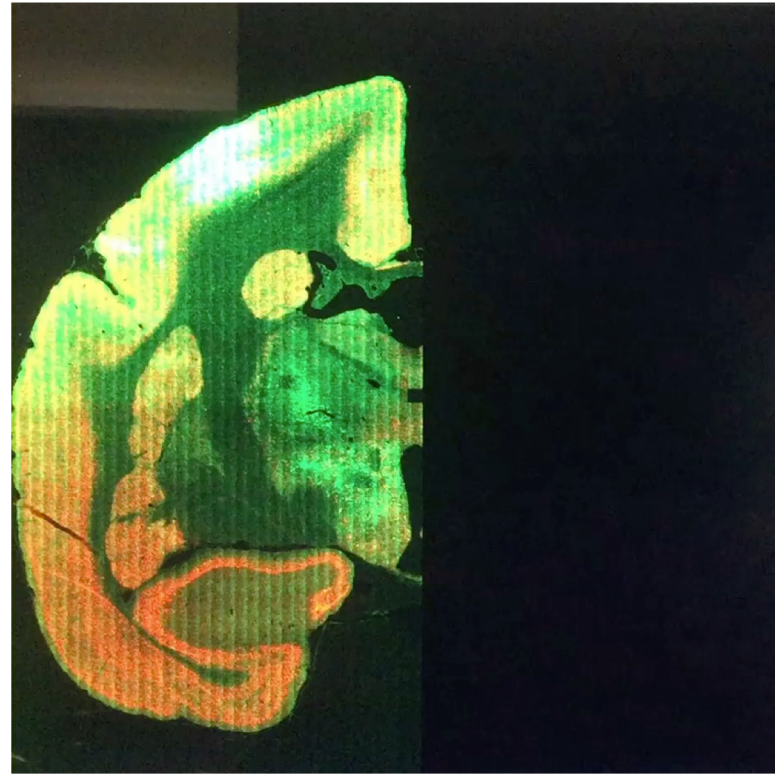
3D Visualization

- Create a national, scalable archival solution to analyze, mine, share and interact with large brain image datasets, collections, and metadata

  - Be a model for future NIH efforts

  - FAIR compliant permanent repository

  - Contain volumetric images of mouse, rat, and other mammals and targeted experimental data (connectivity between cells, spatial transcriptomics)

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh
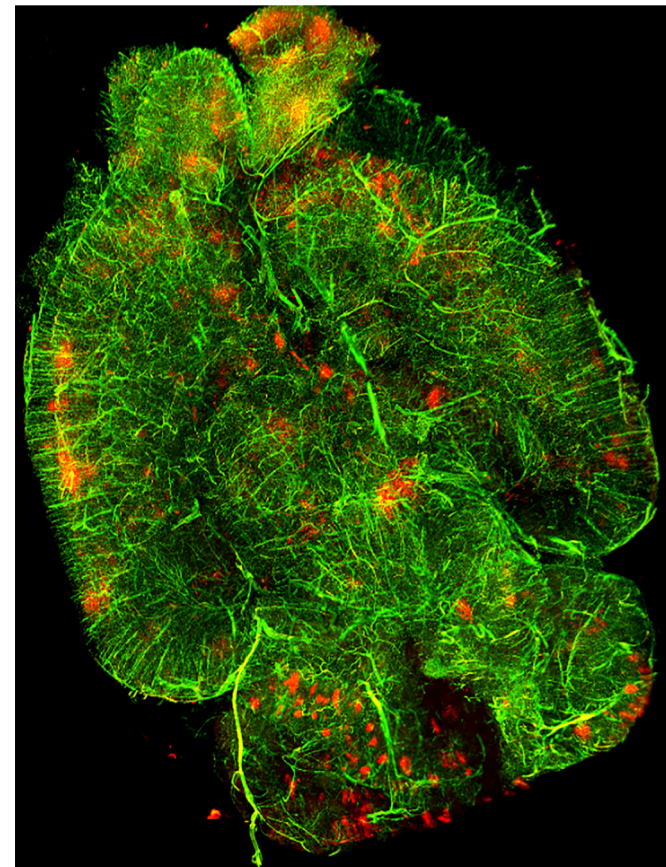
- Provide HPC computing capability local to the data for pre-submission data processing and post-submission exploration

  - Enclave access to pre-release data

  - Research access to restricted-access, secured data

- Provide user access and support (Networking/Data transfer, HelpDesk, Software)


PSC Bridges Virtual Tour:
https://www.psc.edu/index.php/bridges-virtual-tour

Carnegie Mellon University PITTSBURGH SUPERCOMPUTING CENTER University of Pittsburgh

BIL Communities Served

- Data Collaborators/Submitters
  - Archival storage of data
  - Public availability
  - Unique Reference (DOI)
- Data Consumers/End-Users
  - Biomedical Researchers at all experience levels
  - Computer Scientists
  - Other Resources and Databases
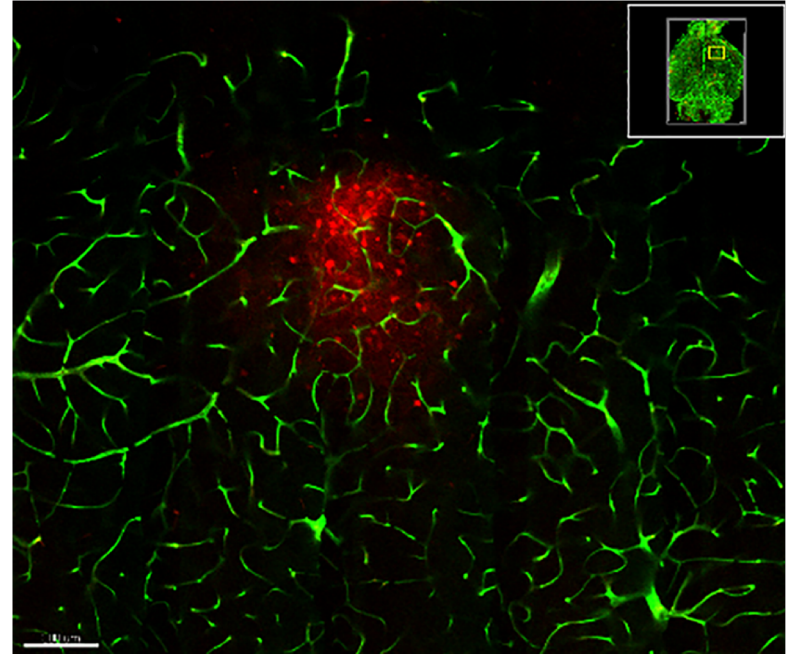
- Remote Visualization:
    - PCoIP (requires special hardware)
    - Virtual Desktop Software w vGPU technology:
        - vGPU technology to enable multiple users to share a physical GPU i.e. desktop runs using 1/8 of a P100 and sends the encoded video to a client.
        - Ex: VMware vSphere, Citrix XenDesktop, and oVirt/KVM
    - WebGL
- Data Management:
    - iRODS (Integrated Rule-Oriented Data System)

BIL Data Management: iRODS

Brain Image Library **iRODS Goals**

- Data and Workflow Management for data collection validation, ingest
- Metadata management and access for query interfaces
- Unified namespace across various distributed storage technologies
- Enable logical view into the data, distributed geographically, and at scale
- Flexible API access to metadata and data collections
- Data replica management and use of tiering to support data backup
- Security and flexible access control for data collections
- Data access logging, audit and measurement automation
- Sustain long term production data management environment

Challenge: Identifying Data Contributors

- Collaboration with BRAIN Initiative Cell Census Network
  - Mini Atlas Project:122 datasets
  - Substantial scale up next fiscal year
- Expect data from other NIH groups next fiscal year

Datasets by Location

■ Allen ■ CSHL ■ USC

Datasets by Project

■ Allen   ■ CSHL_1
■ CSHL_2 ■ CSHL_3
■ Salk   ■ USC
■ Others
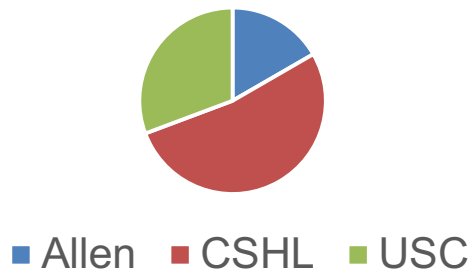
Allen: Allen Institute for Brain Science, Seattle, WA
CSHL: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
USC: Mark and Mary Stevens Neuroimaging and Informatics Institute,
         University of Southern California, Los Angeles, CA
Salk: Salk Institute, La Jolla, CA

Carnegie Mellon University   PITTSBURGH SUPERCOMPUTING CENTER   University of Pittsburgh
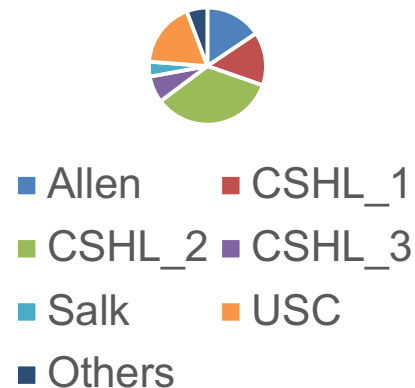
Challenge: Metadata

Metadata Type	Use
Descriptive Metadata	For finding or understanding a resource
Administrative Metadata - Technical - Preservation - Rights	For decoding and rendering files Long-term management of files Intellectual property rights
Structural Metadata	Relationships of parts of resources to one another

From: NISO Understanding Metadata Primer

Challenge: Metadata

Various Metadata
- Instrument settings
- Experiment parameters and conditions
- Subject treatments
- Classification terms
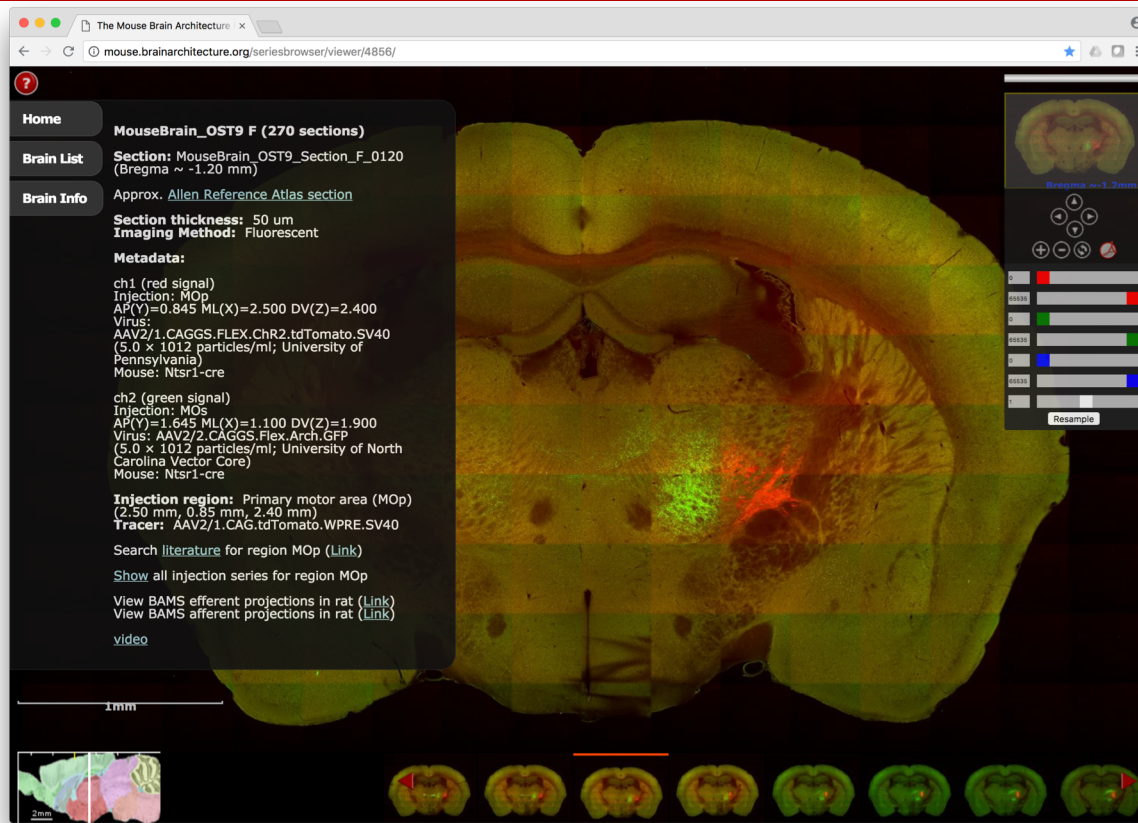- Links to publications
- Links to correlated data
- …

Image from http://mouse.brainarchitecture.org/seriesbrowser/viewer/4856/ Cold Spring Harbor Laboratory

Challenge: Data Movement

- Connectivity to Data Submission Sites:
  - 100Gb/s Internet2
  - 30Mb/s Commodity connectivity
  - PSC Networking Group works with collaborators to optimize throughput between their sites and PSC
- Experimental *BrainBall* Unit:
  - Rugged Disk Enclosure and Shipping Container
  - Low Cost < $2k
  - ~30TB capacity (4x10TB drives in a simple RAID5 setup)
  - Multiple physical port types for connectivity (e.g. USB 3.0, eSATA)

Acknowledgements

- Contributing PSC Staff:
  - BIL Team, Systems & Networking, and Allocations Groups
- National Institutes of Mental Health award R24MH114793
- **iRODS Gurus & Support Team**
  - Especially Jason Coposky & Terrell Russell
- Science Gateways Community Institute
- Brain Initiative Cell Census Network esp. Mike Hawrylycz
- Watson AM, et. al. (2017) Ribbon Scanning confocal for high-speed high-resolution volume imaging of brain. PLOS ONE 12(7):e0180486. (Brain Images and Movie used in this presentation)

Questions?

- Thanks!
    - We look forward to learning from other iRODS installations and users, and to sharing the lessons we learn!

- General questions regarding the Brain Image Library may be directed to bil-support@psc.edu

Carnegie Mellon University

PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh