NASA Data and Computing Architecture Workshop

Philip Blood, Ph.D.
Scientific Director
PSC

Sergiu Sanielevici, Ph.D.
Director of Support for Scientific Applications
PSC

Jonathan Silverstein, MD, MS, FACS, FACMI
Chief Research Informatics Officer
Department of Biomedical Informatics
University of Pittsburgh
PSC: A place for getting research done...

National service provider

Collaborative Research
- $100M research funding

Education and training
- Lead national & local workshops
- Increase diversity in STEM
- Support courses at CMU and elsewhere

PSC IS A JOINT COMPUTATIONAL RESEARCH CENTER BETWEEN CARNEGIE MELLON UNIVERSITY AND THE UNIVERSITY OF PITTSBURGH.

36 years of leadership

23 HPC systems
(11 of which were the first or unique)

Pioneering Artificial Intelligence and Research Computing.

CMU/Pitt Community
- Research collaborations
- Colocation: lower cost and greater reliability & capability

Networking and security
- Networking & security service provider
- Research networking and security

Industry Partners
- Training, access to advanced resources, collaborative research

© Pittsburgh Supercomputing Center, All Rights Reserved
Highly innovative platform for larger and faster artificial intelligence research

Introducing capabilities to accelerate artificial intelligence, machine-learning, and data analytics.

Designed to accelerate discovery for research including:

- Feature recognition in extremely large medical image data sets
- Training of extremely large and extremely challenging problems in astrophysics, weather, genomics and across all research disciplines

For more information: [https://www.cmu.edu/psc/aidb/necortex](https://www.cmu.edu/psc/aidb/necortex)

For more information: [https://www.psc.edu/bridges-2](https://www.psc.edu/bridges-2)

Provides transformative capability for rapidly evolving, computation-intensive and data-intensive research, creating opportunities for collaboration and convergence research.

More Science: Approximately 3x larger than Bridges
Faster Computing: Latest AMD EPYC processors
Faster Storage: Fast flash array
Smarter Science: Designed for AI and data-centric computing
Scalable: Interoperability with cloud and campus resources

For more information: [https://www.psc.edu/anton](https://www.psc.edu/anton)

Accelerating biomolecular simulations

Enables breakthrough research in areas such as medicine, biology, and biochemistry.

Custom-designed hardware for simulating biological molecules at the atomic level with molecular dynamics orders of magnitude faster than standard computers.

Cerebras and HPE are delivering Neocortex

For more information: [https://www.cmu.edu/psc/aidb/necortex](https://www.cmu.edu/psc/aidb/necortex)

For more information: [https://www.psc.edu/anton](https://www.psc.edu/anton)
Bridges-2 High-Level Architecture

Web Server Nodes (6)
- 2× AMD EPYC 7742 CPUs
- 256GB RAM

Database Nodes (12)
- 2× AMD EPYC 7742 CPUs
- 512GB RAM

Data Transfer Nodes (2)
- 4× Intel Xeon Platinum 8160M CPUs
- 4TB RAM

Login Nodes (2)
- 2× Intel Xeon Gold 6248
- 256GB HBM2, 512GB RAM

Users, XSEDE, campuses, instruments

Interconnect

Management Nodes (12)

Tape Backup and Archive
- ~8.6PB usable

Parallel File System
- 15PB usable, 129 GB/s R, 142 GB/s W

Flash Array
- 100TB, 9M IOPs, 100 GB/s

RM Nodes (488)
- 2× AMD EPYC 7742 CPUs, 256GB RAM

LM Nodes (16)
- 2× AMD EPYC 7742 CPUs, 512GB RAM

EM Nodes (4)
- 4× Intel Xeon Platinum 8160M CPUs, 4TB RAM

GPU Nodes (24)
- 8× NVIDIA V100-32 SXM-2, 2× Intel Xeon Gold 6248, 256GB HBM2, 512GB RAM

HPE DMF

Flash Array
- 100TB, 9M IOPs, 100 GB/s

Tape Backup and Archive
- ~8.6PB usable

Parallel File System
- 15PB usable, 129 GB/s R, 142 GB/s W

Interconnect

Management Nodes (12)

GPUDirect

HPE DMF

Flash Array
- 100TB, 9M IOPs, 100 GB/s

Tape Backup and Archive
- ~8.6PB usable

Parallel File System
- 15PB usable, 129 GB/s R, 142 GB/s W

Interconnect
Enabling world-class computational research

Cancer Treatment
- AI to Build Predictive Models of Tailored Treatment
  Joel Saltz & Raj Gupta, Stony Brook University

Breast Cancer Screening
- Deep Learning for Large Volume Cancer Imaging
  Shandong Wu, University of Pittsburgh

Cancer Genomics
- Epigenomic Profiling of Cancer Cells
  David Valle-Garcia, Harvard Medical School

Climate Change
- Identifying Susceptibility to Climate Change
  Rachael Bay, UC Davis

Disaster Prediction
- Numerical modeling of Tsunami Generation, Propagation and Coastal Impact
  Stephen Grill, University of Rhode Island

Astronomy
- Deep Learning for Large-scale Astronomical Surveys
  Asad Kahn et al., University of Illinois

Quantum Materials
- Quantum 2D Layer Optoelectronics
  Pratibha Dev, Howard University

Neuroscience
- Mapping Between Language Models and Brain Activity
  Leila Wehbe, Carnegie Mellon University

Nanomachines
- Protein Unfolding and Translocation by Biological Nanomachines
  George Stan, University of Cincinnati

Energy Storage
- Computational Design of Novel Materials for Energy Storage
  Shyue Ping Ong, UCSD

Behavior Science
- Mouse Behavior Recognition with Deep Neural Networks
  Vivek Kumar, The Jackson Laboratory

Strategic Planning
- Surpassing Human Expertise in Poker
  Tuomas Sandholm & Noam Brown
  Carnegie Mellon University
Bridges-2 has allowed us to speed up our analysis greatly by parallelizing our raytracing code across a handful of nodes. We are particularly grateful to the technical support team that helped us adapt our code to run efficiently at a large scale. In addition to accelerating my research, learning to use Bridges-2 has broadened my understanding of cluster computing. I’m glad to have these concepts in my toolbox as a researcher and programmer.” – Jacob Cohen (UC Davis)

Bridges-2 has been indispensable in facilitating our research. The system is easy to use, and the support teams are always willing to engage and troubleshoot whenever an issue arises (for example, I wanted to run a program last year that would spawn multiple sub-jobs across different nodes, and the Bridges-2 team liaised with me to find a working solution within SLURM). – Saoirse Foley (CMU)

Bridges-2 is powerful and easy to use. Its high memory nodes are critical for certain types of memory-intensive DFT calculations. – Sharon Neufeldt (Montana State University)

Bridges-2, especially the software management, has been extremely helpful. The move to containers makes the service very stable and easy to use. Support of singularity allows us to manage any other packages that would be difficult to install on a cluster otherwise (e.g. Funannotate). - Urs Schmidt-Ott (U. Chicago)

Bridges-2 has been a key resource in enabling our research. The hardware, software and smart/cheerful human support deserves much of the credit for the research we report here. – David Drabold (Ohio University)

Bridges-2 has enabled me to test large numbers of hypotheses about the genomic mechanisms underlying the evolution of mammalian phenotypes. It has also enabled me to test a wide range of ways of training machine learning models for enhancer prediction across species. – Irene Kaplow (CMU)

Bridges-2 allows analysis of datasets larger than previously possible. Seamless transition from high memory jobs to high core jobs allows more efficient processing of data. – Matthew Brian Couger (Brigham and Women's Hospital, Harvard Medical School)
Some best practices as a national service provider

PSC Best Practices

- Flexible queue structure
- Availability for short/immediate jobs as well as long running jobs
- Reservations available for large number of nodes and/or long duration jobs upon request and approval
- Available consultation
- Dedicated assistance for large and time sensitive projects
- Ability to troubleshoot and modify workflow on the fly

Examples

- High Performance Modeling and Analytics for National Scale COVID-19 Response
  - Overnight runs for ~6 months so as not to impact classwork
  - Dedicated staff available for setup and troubleshooting
  - Work informed officials regarding best practices and steps
- Louisiana Compound Flood Forecasting System Development
  - Staff ready to spin up reservations with a 12 hour lead time
  - Real time Hurricane monitoring
# Improving Our Processes

## Continuous Improvement Committee (CIC)

Founded on March 6, 2021

**77** suggestions received as of September 1, 2022.

**58** tasks have been completed.

*Examples*: Retaining job submission files, OnDemand enhancements, revising error messages, system monitoring tools, documentation clarifications, modifications to Filemover software to rapidly migrate data between filesystem tiers.

## Task Forces (SOP)

**Software Installation**

Developed and published a unified processes for installing, deploying, creating modules for, and documenting software installed on B2 to provide further clarity to researchers leveraging these tools.

**Services**

Define the services (VMs, web servers, databases, etc.) in scope for B2; document how to request and support said services for our community.

## Monitoring / Communicating

**Dashboards**

Bring data to the forefront of our decision making by graphing utilization across partitions, modules, groups, FOS. Monitoring data center environments and system critical infrastructure.

**Remote Communications**

Exploring ways of providing quality support to our communities in the platforms they're actively leveraging (Slack, Discourse, etc.)
XSEDE: A national cyberinfrastructure

XSEDE COVID-19 UPDATES

WHAT IS XSEDE?

The Extreme Science and Engineering Discovery Environment (XSEDE) is a single virtual system that scientists can use to interactively share computing resources, data and expertise. People around the world use these resources and services — things like supercomputers, collections of data and new tools — to improve our planet.
XSEDE Features Designed to Enhance User Productivity

- Unified XSEDE User Portal
  - coordinate and unify services and information
  - training offerings, allocations, documentation, publication management, resource status and performance
- Unified allocation of resources
  - ensure that most meritorious work nationally is awarded the most suitable resources
  - ease of transfer of allocations to provide flexibility for changing research needs
- Single Help Desk
  - single user interface to manage/track tickets
  - users have one point of contact for all issues
- Coordinated advanced support and training
- Application software catalog and integrated resources catalog
XSEDE Features Designed to Enhance User Productivity (2)

• Unified authentication mechanisms
  • option to use local credentials

• Coordinated security
  • has prevented the spread of security breaches from site to site

• Common techniques for rapid data transfer (Globus)

• Support for end-to-end network tuning to optimize data transfer

• Well-defined infrastructure
  • allows campuses and regional infrastructures to design their systems to interoperate with those of XSEDE
  • attracts new service providers (4 providers allocating resources via XSEDE for the first time in 2021; more to come)
Characteristics of Examples of Science Impact

- Use of resources at multiple sites
  - facilitated by XSEDE’s unified allocations
- Extended Collaborative Support Service
  - improves the efficiency of codes
  - implements best-practice workflows and data management
  - makes certain calculations feasible where they were not before
- Code improvements developed with a single group have made their way back into community code releases
  - improve productivity for entire communities
- Engage new communities by the proactive efforts of the Novel and Innovative Projects team
  - machine learning, history, social sciences...
Mission: Improve the productivity of the XSEDE user community through successful, meaningful collaborations to optimize their applications, improve their work and data flows, and increase their effective use of the XSEDE digital infrastructure and broadly expands the XSEDE user base by engaging members of underrepresented communities and domain areas.
ECSS Areas of Focus

• ECSS consists of five areas. End users don’t need to be aware of the areas and our staff will figure out where a project fits best:
  • Extended Support for Research Teams (ESRT)
    • “traditional” ECSS projects to improve software and make best use of XSEDE resources
  • Extended Support for Community Codes (ESCC)
    • similar to ESRT, but with emphasis on codes available for public use
  • Extended Support for Science Gateways (ESSGW)
    • application of technologies that enable access to HPC through web interfaces
  • Novel and Innovative Projects (NaIP)
    • assistance to users from domains that are relatively new to XSEDE and high performance computing
  • Extended Support for Training, Education and Outreach (ESTEO)
    • training and technical support for use of advanced cyberinfrastructure
ECSS as an allocated service

• Expert staff can be requested for collaborations lasting months to a year
  • Expert staff can be requested for collaborations lasting months to a year

• Typical collaborations require 20-25% staff time for one year

• Critical mass engenders success
  • 31 FTEs (70 individuals) at ten XSEDE sites
  • ECSS staff have advanced degrees in a variety of science, computing and technology fields
  • some staff co-author publications or write proposals with PI team, a few are later funded by PI team
Ten Years of Collaboration Yields Transformative Outcomes in Genomics

• In 2011, three researchers reached out to XSEDE for help with large-scale, cutting-edge genomics work
  • Brian Couger then a graduate student at Oklahoma State
  • Chris Mason then an assistant professor at Cornell, and his graduate student, Lenore Pipes

• Researchers needed three things
  • large memory systems accessible through XSEDE
  • help running genomics software at scale
  • help orchestrating big data genomics workflows on XSEDE systems
XSEDE ECSS Catalyzes Knowledge Transfer

- Through extensive collaboration with XSEDE ECSS, Couger and Pipes developed deep experience running large-scale genomics workflows on national cyberinfrastructure.

- In 2013-2014, these researchers co-authored papers on best practices with ECSS.

- Couger brought this experience to many research collaborations, and his new position at Harvard.

- Lenore Pipes transferred her experience to others in the growing Mason lab at Cornell, and current lab at Berkeley.
Using advanced large-scale genomics workflows, Couger determined that the male creeping voles have two X chromosomes. The way this rodent species determines sex differs from all other mammals known. These results solve a 60-year-old mystery in biology and raise important questions on how early sexual development works.
Massive Collaborations on Biology of Spaceflight and Global Map of Urban Microbes

- The Mason lab assembled genetic sequences from the DNA of two identical-twin astronauts and compare these to the data from 59 other astronauts, showing increased mitochondrial mutation due to spaceflight
- Mason’s lab also headed up an international collaboration of scientists to create a global map of urban microbes and antimicrobial resistance
NSF ACCESS Program Structure

ACCESS Services

Allocations
- Allocation Services
- Innovative Pilots
- Service Model

Support
- General User Assistance
- Allocation & Utilization Assistance
- End User Training
- Computational Science Support Network

Operations
- Operational Support
- Data & Networking Support
- Cybersecurity Support

Metrics
- MDM Operations
- Service Model
- Data Analytics Framework

ACCESS Coordination Office Services

Executive Council

Communications & Outreach

Advisory Board
Standardized Allocation Opportunities

Four Allocation Opportunities to suit a variety of needs (core hours):

- **Explore (400,000)**
  - Best-suited for endeavors with light resource requirements
    - Grad students can be PIs
- **Discover (1,500,000)**
  - Minimal effort to start production research activities
    - Potential best-fit for Campus Champion Allocations
- **Accelerate (3,000,000)**
  - More substantial resource requirements
    - Multi-grant research, Gateways, etc.
- **Maximize (No upper limit)**
  - For large-scale research projects with extreme resource needs

Quick FAQs:

- Except for Maximize, opportunities are available for request at any time.
  - Maximize will follow the XRAC schedule and policies until March 2023 meeting.
    - Bi-annual opportunities will then be available.
- Multiple allocations can be requested
  - For multiple grants
  - Or for substantially different needs
Innovative Pilots for Allocating Cyberinfrastructure

Longer-term vision for novel projects that allow researchers even greater access to CI resources and capability.

● **Campus On-Ramps**
  ○ Will “place” the proposal submission interface directly into University portals
    ■ To engage with researchers in their “comfort zone” and provide institutions with a virtual datacenter of NSF-funded resources

● **Allocating CI Workflows**
  ○ Allocating resources as part of constructed workflows to build workflows from edge resources to a “central” resource
    ■ Will incorporate a range of resource types — from instruments to HPC

● **Sensornets and Instrumentation**
  ○ Calendar-level scheduling for resources that can’t be shared simultaneously (telescopes, aircraft, etc.)
    ■ Will be able to bundle resources for data storage and processing
Leading data infrastructure

Developing the tools to create an open, global atlas of the human body at the cellular level.

18 Research Group Consortium Funded through NIH Common Fund

For more info: http://hubmapconsortium.org

Brain Imaging Library

Creating a national, public resource to deposit, analyze, min, share and interact with large brain imaging datasets.

Alex Ropelewski
Director of Biomedical Applications
Pittsburgh Supercomputing Center

Alan Watson
Assistant Professor
University of Pittsburgh

For more info: http://brainimagelibrary.org

Other research projects include:

- Open Compass: Academic advanced engineering testbed for AI
- HPN-SSH: Accelerating highly secure data transfer
- NCGAS: National Center for Genome Analysis Support
- MMBios: National Center for Multiscale Modeling of Biological Systems

PSC has received generous support from:

- National Science Foundation
- National Institutes of Health
- The Commonwealth of Pennsylvania
- The Bill and Melinda Gates Foundation
- Department of Defense

Carnegie Mellon University
Pittsburgh Supercomputing Center
University of Pittsburgh
HuBMAP is working to catalyze the development of a framework for mapping the human body at single cell resolution.
HuBMAP Goals

**Accelerate Tools and Technique Development**
Accelerating the development of the next generation of tools and techniques for constructing high resolution spatial tissue maps that quantify multiple types of biomolecules either sequentially or simultaneously.

**Generate 3D Human Tissue Maps**
Generating foundational 3D human tissue maps using validated high-content, high-throughput imaging and omics assays.

**Establish Open Data Platform**
Establishing an open data platform that will develop novel approaches to integrating, visualizing and modelling imaging and omics data to build multi-dimensional tissue maps, and making data rapidly findable, accessible, interoperable, and reusable by the global research community.

**Collaborate with the Research Community**
Coordinating and collaborating with other funding agencies, programs, and the biomedical research community to build the framework and tools for mapping the human body at single cell resolution.

**Support Pilot Projects**
Supporting pilot projects that demonstrate the value of the resources developed by the program to study normal individual variations and tissue changes across the lifespan and the health-disease continuum.
Unique features of HuBMAP

- Free, open access to non-identifiable data, tools and technologies developed
- Healthy tissue
  - Diversity in samples: sex, ethnicity, age
- Comprehensive accessible three-dimensional molecular and cellular atlas of the human body
- Culture of openness and sharing using team science-based approaches
  - Protocols.io, github
- Diverse expertise in funded members
  - Molecular, cellular, developmental, and computational biologists, measurement experts, clinicians, pathologists, and software engineers, and computer and data information scientists
- Coupled with leading HPC resources and cloud burstability.
The Consortium

- Rapid Technology Implementation
- National Institutes of Health
- Tissue Technology Development
- Tissue Mapping Centers
- HuBMAP Integration Visualization Engagement
Contributing Sites for Ramp Up Phase (years 1-4)

Where are they!?
The Human Body at Cellular Resolution: The NIH Human Biomolecular Atlas Program.
Establishing a FAIR, open, and flexible data platform

Findable, Accessible, Interoperable, Reusable (FAIR)
- Open software, data, and publication policies
- Flexible hybrid cloud microservices architecture
- Robust provenance with unique identifiers
- Ontologies and standardized metadata
- Portable/interoperable/reusable/reproducible pipelines
- Community standards for publishing findable and accessible APIs and data
HuBMAP Policies

- Associate Member Guidelines
- Consent Policy
- Data Sharing Policy
- Data Use Agreement
- Material Transfer Agreement
- NIH Genomic Data Sharing Policy
- Publication Policy
HuBMAP Hybrid Cloud Microservices Infrastructure

https://software.docs.hubmapconsortium.org/hubmap-services-arch.html
Human BioMolecular Atlas Program
An open, global atlas of the human body at the cellular level

The HuBMAP Data Portal is the central resource for discovery, visualization, and download of single-cell tissue data generated by the consortium. A standardized data curation and processing workflow ensure that only high quality is released.

Explore spatial single-cell data with Vitessce visualizations

View multi-modal assay types with reusable interactive components such as a scatterplot, spatial imaging plot, genome browser tracks, statistical plots and controller components.

Get Started

https://portal.hubmapconsortium.org
HuBMAP Data Portal
portal.hubmapconsortium.org

Overview of Data Portal Tools

**Common Coordinate Framework (CCF)**

Kidney and spleen tissue sample locations are mapped 3-dimensionally against a human reference image.

**Azimuth**

Explore this web application that uses an annotated reference dataset to automate the processing, analysis, and interpretation of a new single-cell RNA-seq experiment.

**Vitessce Image Viewer**

Enables the exploration of spatially-resolved, integrated single-cell datasets.

**Extension User Interface (EUI)**

Supports exploration of semantically and spatially explicit data - from the whole body to the single cell level.

**Registration User Interface (RUI)**

Supports uniform tissue data registration across organs and labs.

**ASCT+B Tables**

Anatomical Structures, Cell Types, plus Biomarkers (ASCT+B) tables aim to capture the nested part-of structure of anatomical human body parts, the typology of cells, and biomarkers used to identify cell types.
Current state of Submitted, Processed, and Published Data

Interactive Sankey Diagram
HuBMAP Data Ingestion, Processing, and Publication
Provenance, Metadata and Data Collection

HuBMAP uses a highly flexible provenance Model based upon Prov W3C Standard.

Entities must be registered before uploading data.

Business logic requires pattern of provenance pictured.

- Donor Data
  - Demographics information
  - Clinical data
  - Selection Criteria
- Tissue Sample Metadata
  - Sample prep details
  - Sample prep protocols
  - Location information within organ
- Assay Metadata
  - Instrument details
  - Experiment details
- Centrally processed datasets (Assay Data)
HuBMAP Prov schema based upon W3C PROV
Registering Tissue Samples

From an organ, multiple levels of samples can be registered. In this case, we’re registering 100 tissue sections from a block of tissue, so 100 new HuBMAP IDs are generated.
The spatial location and anatomic semantics of a tissue block is collected during the registration flow of the sample. In this case the organ is a kidney so the kidney spatial registration user interface is used to place the block in a kidney.
Registering a Dataset

When a dataset is registered, a directory is created where the data can be uploaded via Globus (below shows a link to go directly to Globus Transfer app to with permission to upload there).

<table>
<thead>
<tr>
<th>Source(s)</th>
<th>Submission ID</th>
<th>Subtype</th>
<th>Group Name</th>
<th>Status</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBM448.NSSK.723</td>
<td>UFL0004-TH-2.4-1</td>
<td>FFPE slide</td>
<td>University of Florida TMC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBM254.RZFN.624</td>
<td>UFL0005-TH-1.1-1</td>
<td>FFPE slide</td>
<td>University of Florida TMC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lab Name or ID

X93-IMS-735

Description

MALDI IMS output from heart sections

Gene Sequences

- No
- Yes

Will this data contain any human genomic sequence data?

Data Type

- MALDI IMS

Save was successful

HubMAP ID: HBM234.MLNW.894
Type: Dataset

Click here to go to the Globus data repository

Done
1. Go to **ingest validation tools**
2. Access **Donor**, **Organ**, **Sample**, **Antibodies**, **Contributors** (at bottom) for .TSV template
3. Download & fill out .TSV template for each relevant domain above
4. Upload .TSVs **as part of submission**
Unified Medical Language System (UMLS)
Over 200 source ontologies

Jonathan Silverstein (HuBMAP) presented a neo4j graph database model/schema based upon UMLS @CFDE meeting as HuBMAP’s approach to Ontologies/Knowledge Management/APIs for general and application ontologies

Knowledge Graph (KG) model of UMLS

Uberon and Cell Ontology were also integrated in for HuBMAP’s use

UMLS is generally built around human concepts

But for this use, we needed to integrate mouse ontologies to support translation of phenotypes and genes.
Extend the KG to integrate ontologies & data!

- genomic resources
- proteomics
- mouse phenotypes
- patient phenotypes
- WGS germline
- WGS tumor
- WES tumor
- radiology
- RNA-seq
- pathology
- epigenetics
- Drugs/perturbagen

**TO DO**

- diseases
  - mondo 30k
- phenotypes
  - hpo 17k
  - uberon
  - anatomy
- gene function & subcell. anatomy
- mouse phenotypes

- KF GTeX
- KOMP2
- All
- GTEx

NCIT 144k

Added Dec 2020

Deanne Taylor
HuBMAP Hybrid Cloud Microservices Infrastructure

https://software.docs.hubmapconsortium.org/hubmap-services-arch.html
Ingestion and Processing

- Containerized tools
- Composed as Common Workflow Language (CWL) workflows
- Run as DAGs in Apache Airflow
- Each step tracked with provenance, metadata, versioning

1. **Data Analysis DAG: Create new derived dataset**
   - Send parent dataset UUID to Ingest API
   - Get UUID and create directory for derived dataset
   - Set status for derived dataset to “Processing”

2. **DAG: Run CWL Pipeline** (provided by CMU-TC)
   - Prepare CWL workflow
   - Execute workflow
   - Verify successful execution
   - Move derived data to UUID directory

3. **DAG: Report Derived Dataset Metadata to Ingest API**
   - File information
   - DAG provenance
   - Change dataset status to “QA”

4. **DAG: Spawn a CWL pipeline task to prepare derived datasets for Data Portal** (provided by Harvard-TC)
   - Repeat steps 1-3
Visualized data after processing
Building a Flexible Analysis Ecosystem

- Motivation: How do we enable scientists to make the most of HuBMAP data and reduce time to science?
- Vision: Build a flexible analysis ecosystem
  - A collection of data, tools, and computational resources co-localized and designed to reduce time to science by reducing user friction.
  - A flexible analysis ecosystem will support highly customizable workspaces where scientists can bring together the open resources and tools they need to do their work.
- First steps
  - With TC-Harvard building support for HuBMAP portal features that grant users the ability to create workspaces and perform custom analyses and visualizations using Jupyter Notebooks.
  - Building a flexible backend to support this use case and eventually the broader vision of a flexible analysis ecosystem.
  - Incorporating a modular approach will allow other teams to deploy this system for their own data portals.

Credit: Juan Puerto, PSC
Flexible Analysis Ecosystem: Modularity

- **Key design goals**
  - Support diverse range of tools
  - Support diverse of storage and compute resources
  - Build a service that is easily deployed at other sites
    - Fully public cloud based deployment
    - Fully On-Premises deployment
    - Hybrid deployment!

- **Reusable modules for different deployment scenarios**
  - Compute Resources (Yellow)
    - How are interactive tools/analytic pipelines run?
  - Storage Resources (Red)
    - Where are results/inputs stored?
  - User Authentication/Authorization (Green)
    - Are users allowed to access these modules?
Flexible Analysis Ecosystem: Example

● Initial backend deployment:
  ○ API Authentication Method (Globus):
    ■ How users are authenticated with the API
  ○ Main Storage Method (HIVE local filesystem):
    ■ Where files are saved by default
  ○ Available Resources:
    ■ PSC HIVE Compute (Slurm)
    ■ PSC Bridges-2 Compute (Slurm)
    ■ AWS EC2 (future work)
  ○ Available Storage:
    ■ HIVE local filesystem (running on HIVE)
    ■ Bridges-2 filesystem (running on Bridges-2)
    ■ S3 (future work)
  ○ Resource Authentication Methods:
    ■ How users are authenticated against specific resources
      ● PSC User API (for both Slurm resources)
      ● Globus (for AWS resource)
HuBMAP Hybrid Cloud Microservices Infrastructure

https://software.docs.hubmapconsortium.org/hubmap-services-arch.html
HuBMAP Data to Common Fund Data Ecosystem (CFDE)

Developing the HuBMAP Data Manifest
- Metadata harmonization
- Handling sensitive metadata
- Push vs. Pull
- Automation
- ...

Details (examples)
- HuBMAP data entities & hierarchies
- Diversity of assay types
- Evolving metadata definitions
- Controlled vocabularies
- Reharmonization
- ...
Single sign-on and data transfer: CFDE and NIH

NIH Research Auth Service (RAS)
GA4GH Data Repository Service (DRS)
Increasing FAIRness

**HuBMAP Public Resources** so far (appearing this month!):

- **52 Repositories** in GitHub (most are Docker tools)...with versioning
- **42 API endpoints** in SmartAPI
- **Hundreds of DOIs** (datasets, collections, and use of protocols.io)
- **Ontologies** in use include UBERON, Cell Ontology, Concepts in UMLS cross-referenced to SNOMEDCT_US, HPO, etc...with versioning

**Website, Data Portal and APIs** use Globus for auth and data

We will use the **FAIRshake CFDE rubric** and other applicable approaches as they mature to iteratively assess and improve HuBMAP FAIRness (portals and data, including via CFDE) to community standards

We will **actively collaborate in CFDE-CC harmonization** and provide as APIs for other DCCs our own developments and approaches to achieve HuBMAP harmonization

We will support CFDE deliverables to FIND, ACCESS, INTEROPERATE, REUSE **team experience** with CEDAR and GO FAIR
Connect with HuBMAP

info@hubmapconsortium.org

www.hubmapconsortium.org

@_hubmap