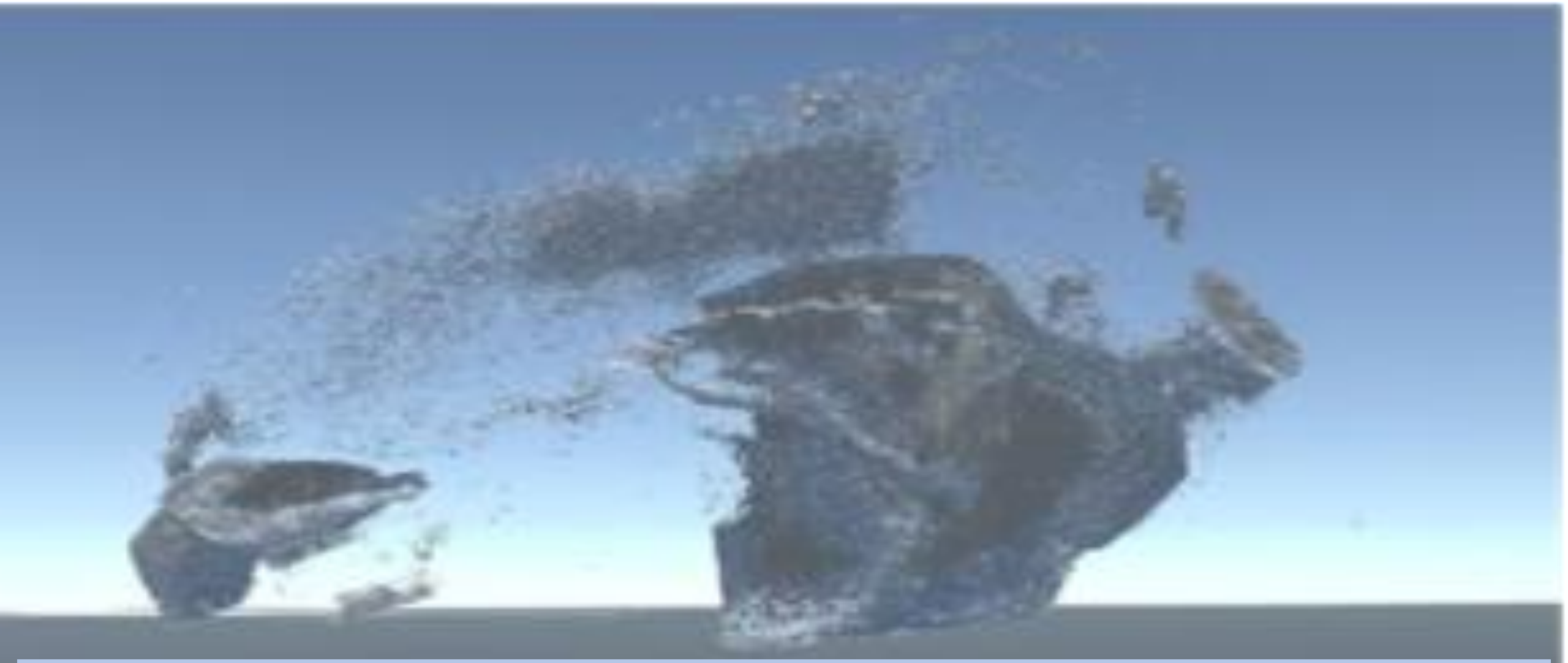
iRODS Impact on Science and Data Management

iRODS UGM 2017

Ashok Krishnamurthy ,Kira Bradford, Michael Conway, Michael Shoffner, Justin James

iRODS impact on data management for Scienctific domains: 2 Use Cases

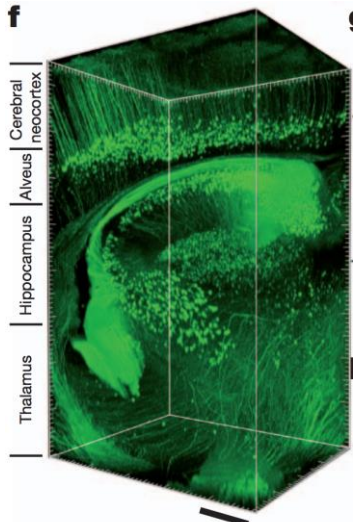
- **BRAIN-I**
  - A unified computation framework for analysis, storage, and visualization of 3D microscopy data of the brain
- **SC2I**
  - Clinical decision support tools to improve medical outcomes in acute care



BRAIN-I: A unified computational framework for analysis, storage, and visualization of 3D brain microscopy data

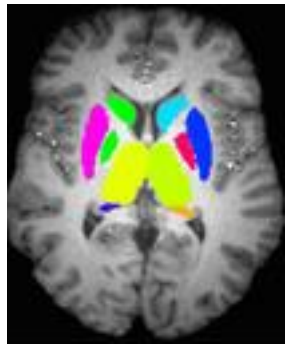
Big Data Problems in Neuroscience

Examples of Big Neuroscience Data



(Chung et al., Nature, 2013)

3D microscopy data (including functional imaging/structural imaging)

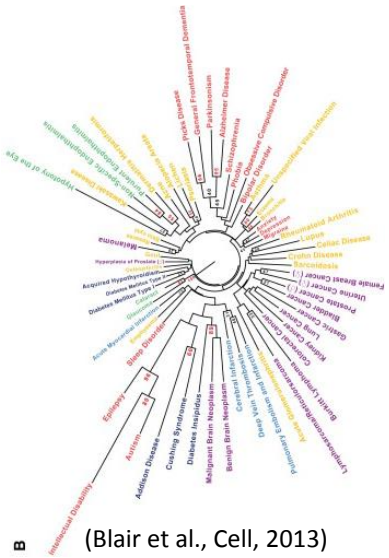


(Hibar et al., Nature, 2015)

Human brain imaging (MEG/EEG/MRI)

(Bras et al., Nature Reviews Genetics, 2012)

Sequencing/genomic platforms (e.g. human whole genome-sequencing, single-cell transcriptomics)



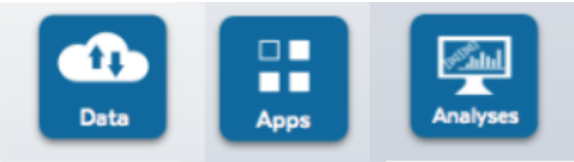
(Blair et al., Cell, 2013)

Electronic Medical Records

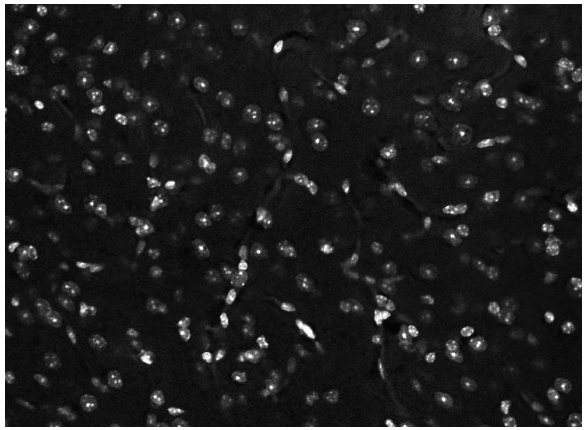
Big data problems

- Sharing & Moving Data
- Searching data within and across labs
- Where to perform large-scale computation
- Making models of brain function
- Visualization of complex data
- Confidentiality of human data

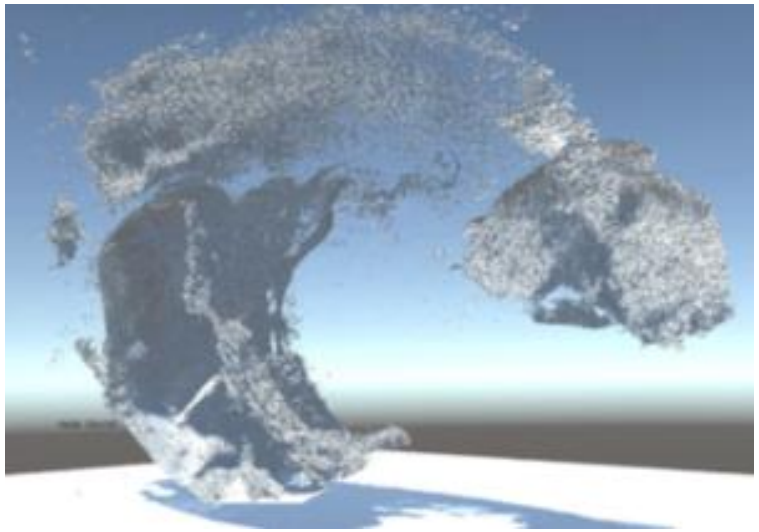
Computational infrastructure for storage, sharing and analysis of 3D microscopy images



Novel segmentation tools to trace brain structure

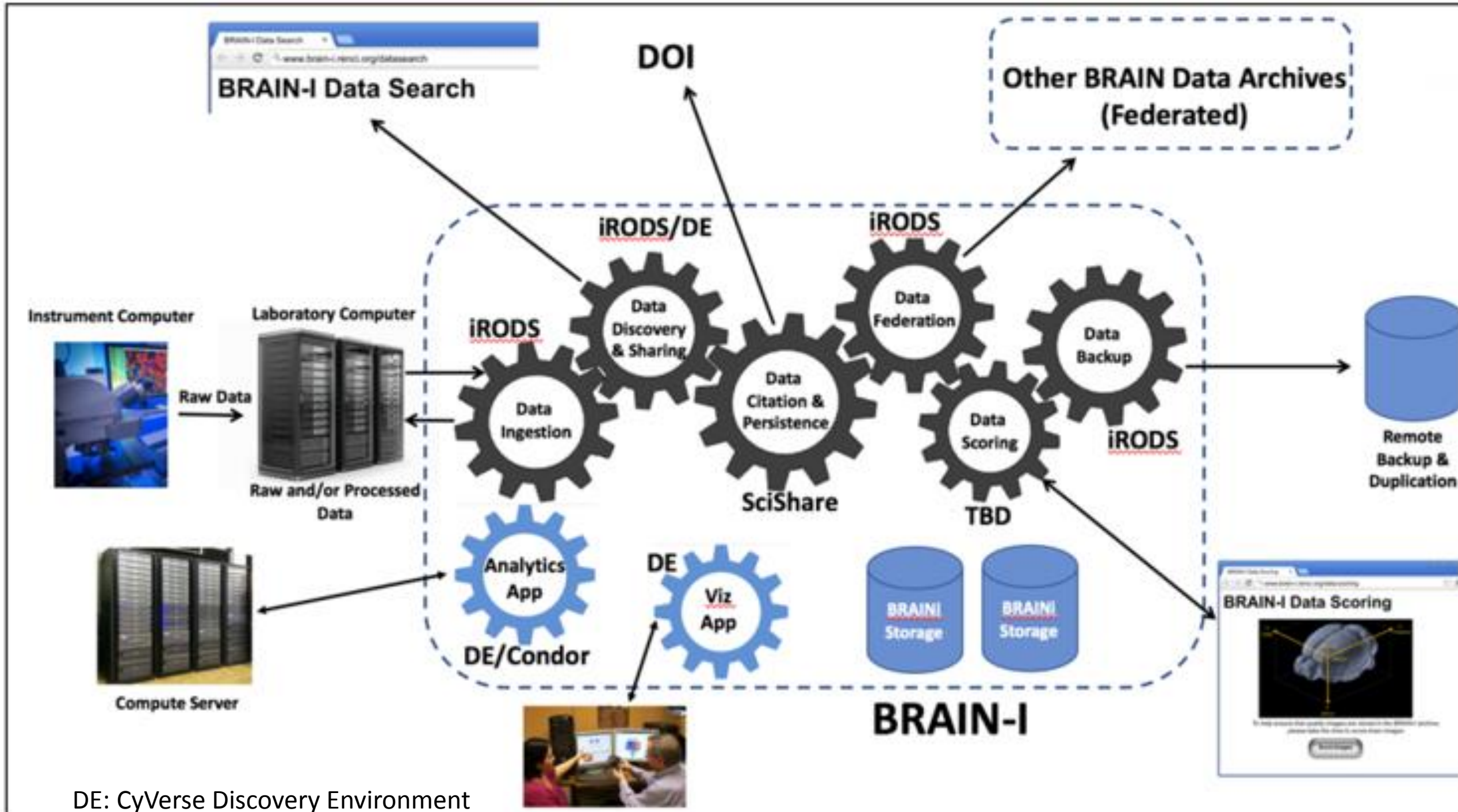


Visualization of 3D brain images using immersive environments



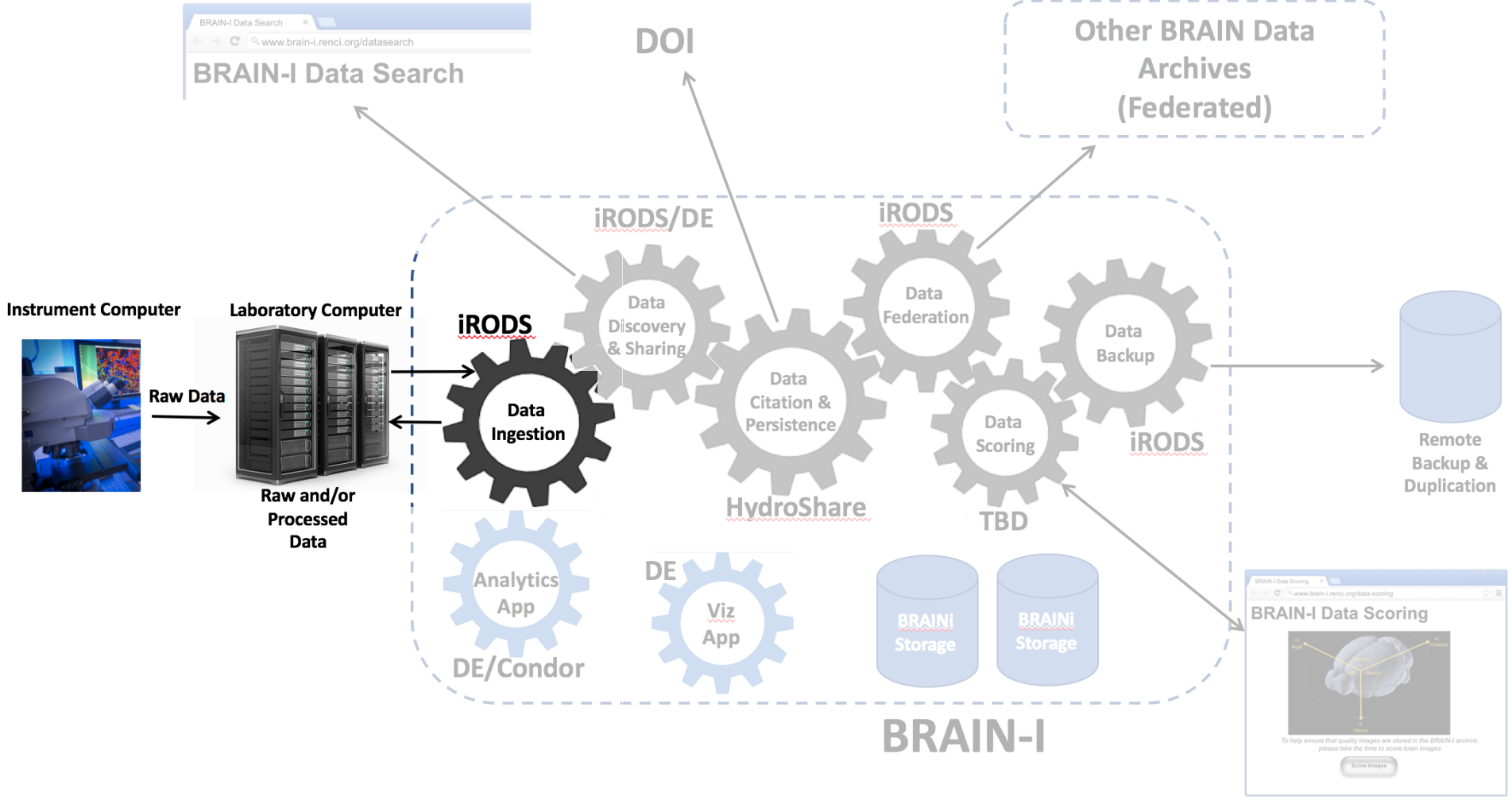
BRAIN-I

Funded by the National Science Foundation

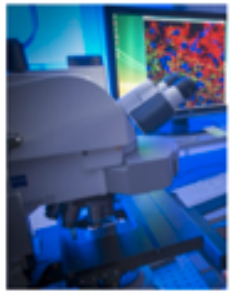


DE: CyVerse Discovery Environment

Data Ingestion



Data Accession Sequence

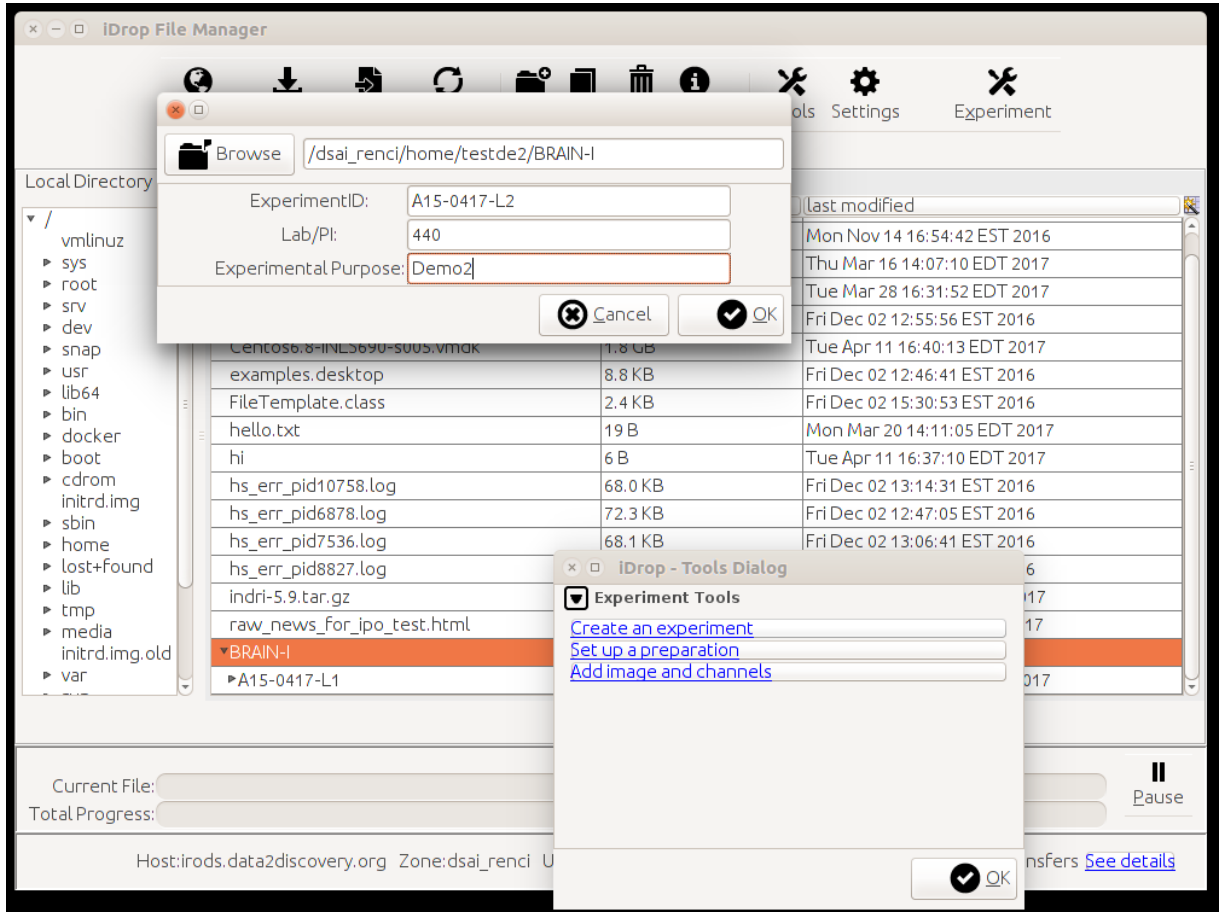


Microscope data and gathered metadata transferred to grid

Validation, Automated extraction of additional metadata via policies and rules

Automated replication of data to BRAIN-I

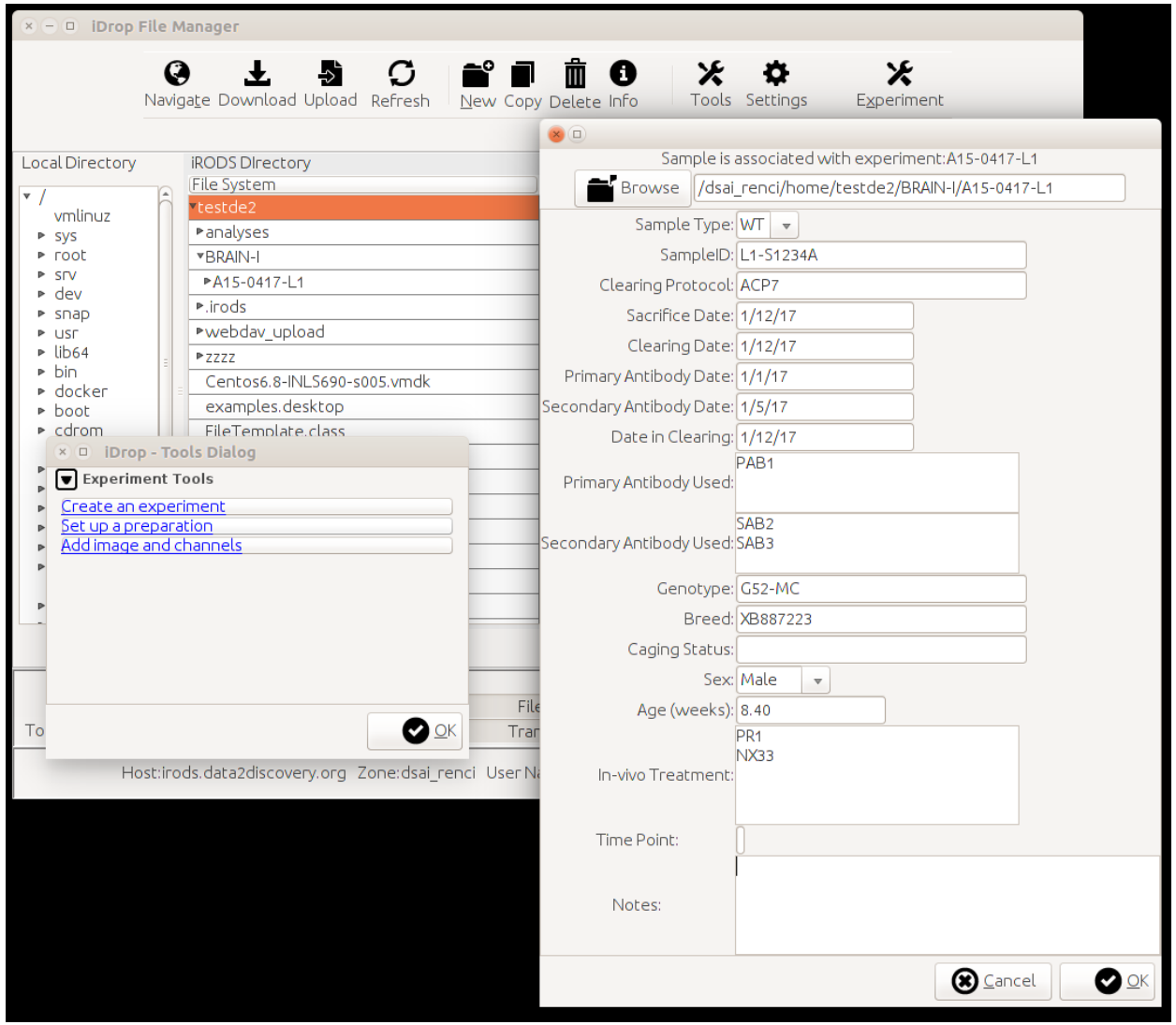
Data Ingestion – Standards and Identifiers



Data Capture on Instrument

- Desktop 'agent' that can manage accession of instrument data to the lab data grid
- Provision metadata for experiments via templates
- Interrogation of instrument for additional metadata

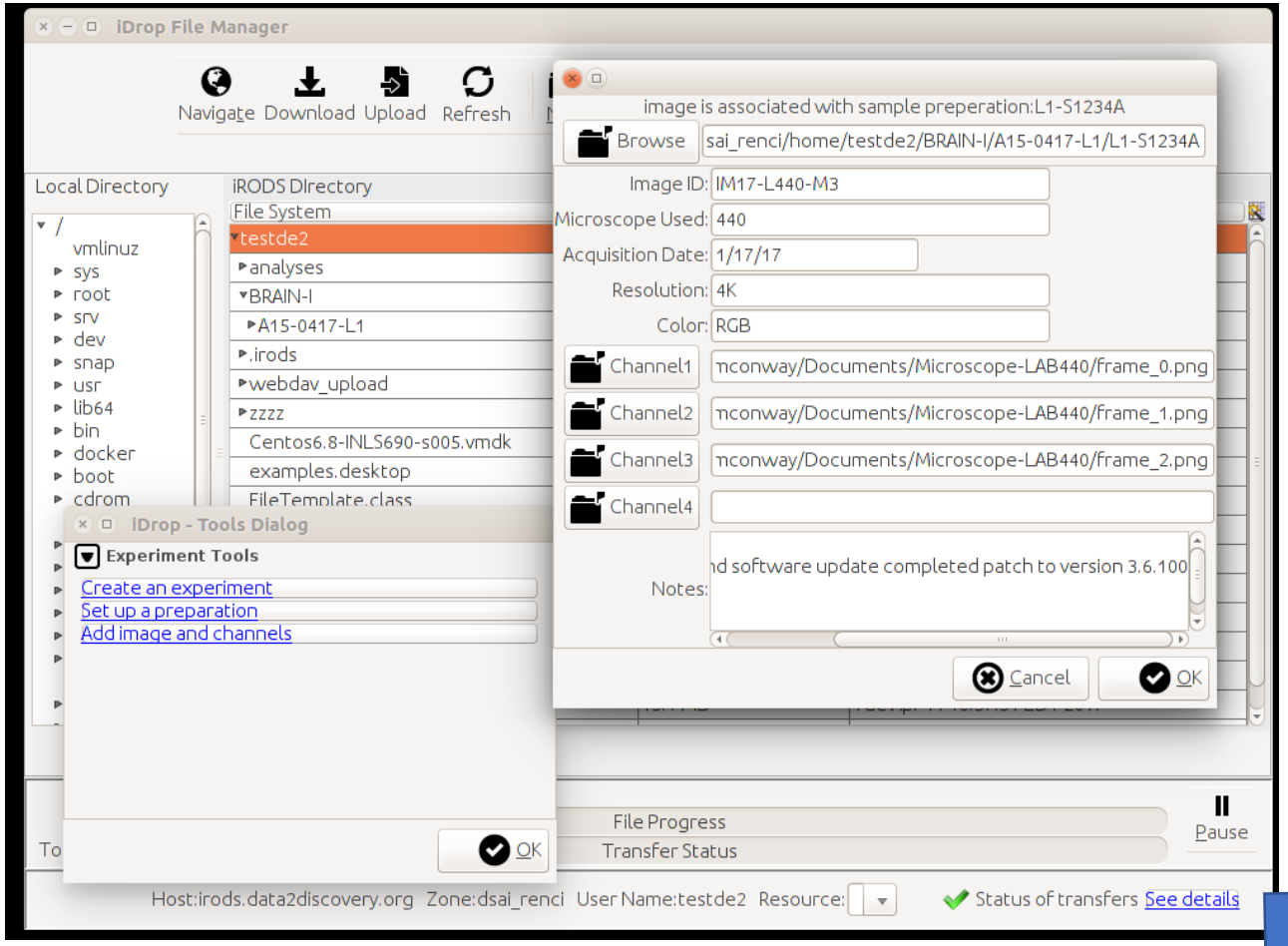
Data Ingestion – Standards and Identifiers



Data Capture on Instrument

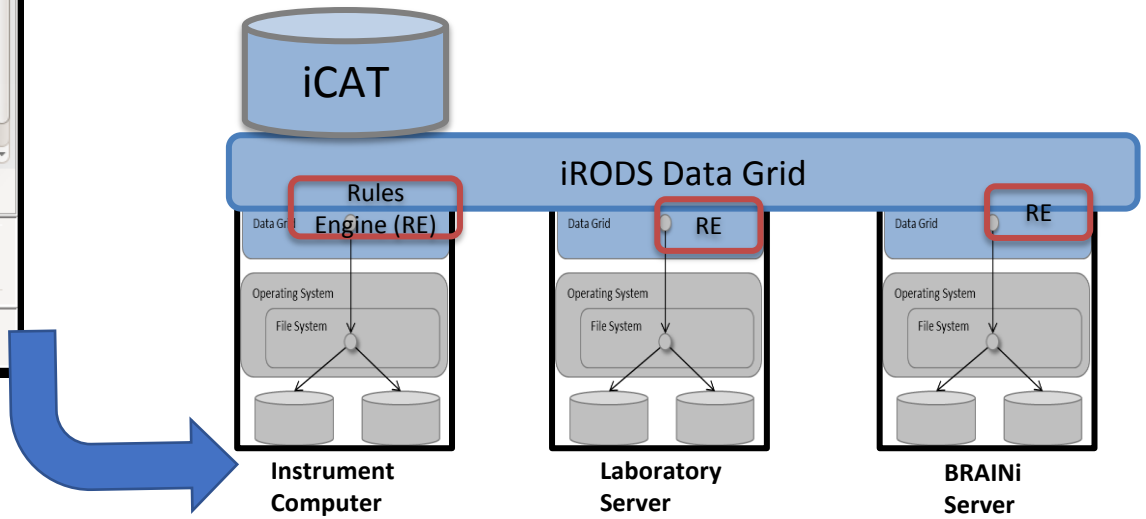
- Adding a prepared test specimen to the experiment
- Common metadata is populated automatically from the template

Data Ingestion – Standards and Identifiers

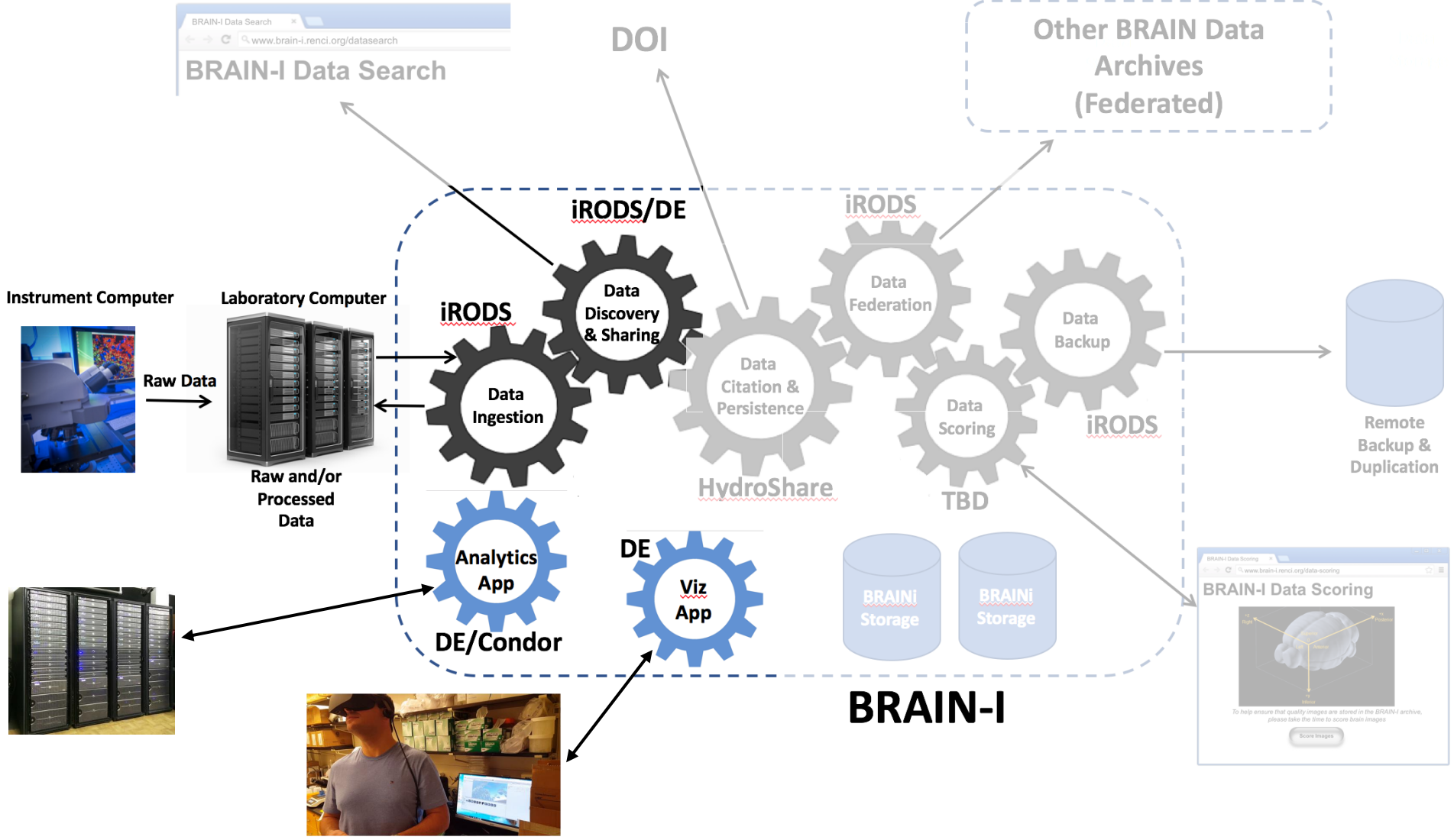


Reliable (hands off) accessioning of curated instrument data

- Image channels identified and linked to sample
- Reliable, auditable accessioning of large files to lab data grid
- Error tracking, reliability
- Ability to schedule multiple accession actions to run overnight

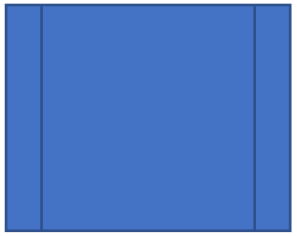


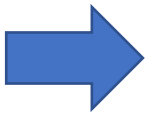
Analysis and Visualization Tools

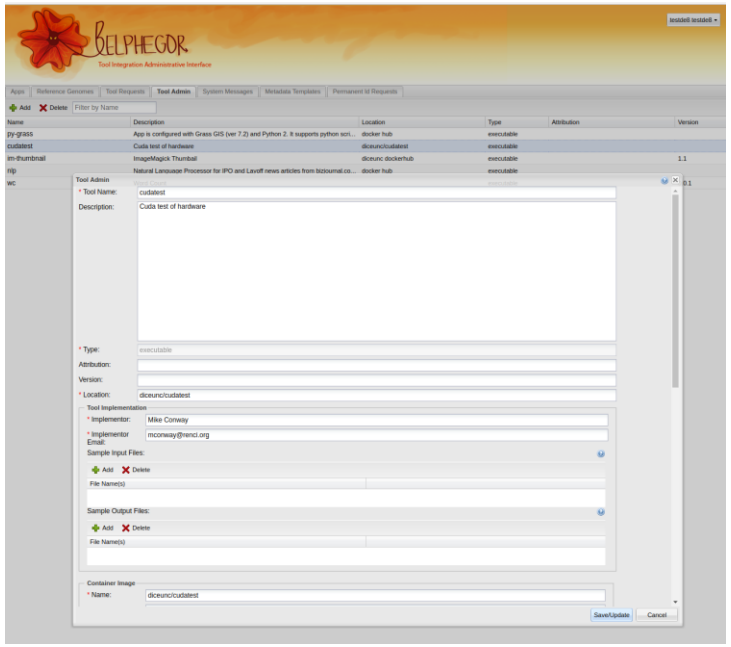


Analysis and Visualization Tools

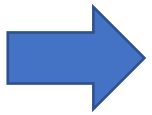
Package any app or algorithm as a Docker image

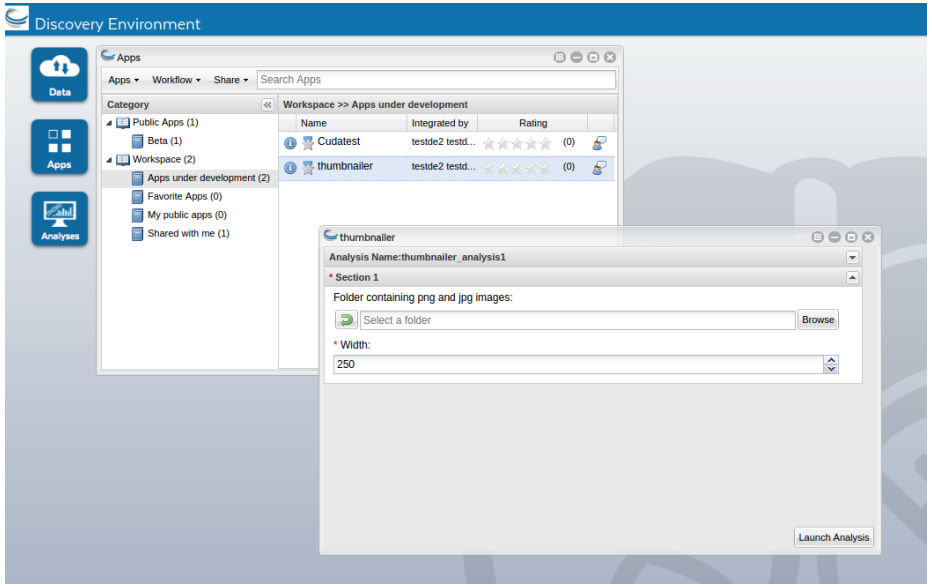






Have an administrator add the app as a 'Tool'





Users can create a GUI to launch the tool, and share these GUI Apps with others

Data replicated to GPU compute resource

Dockerized analysis routed to GPU machine automatically

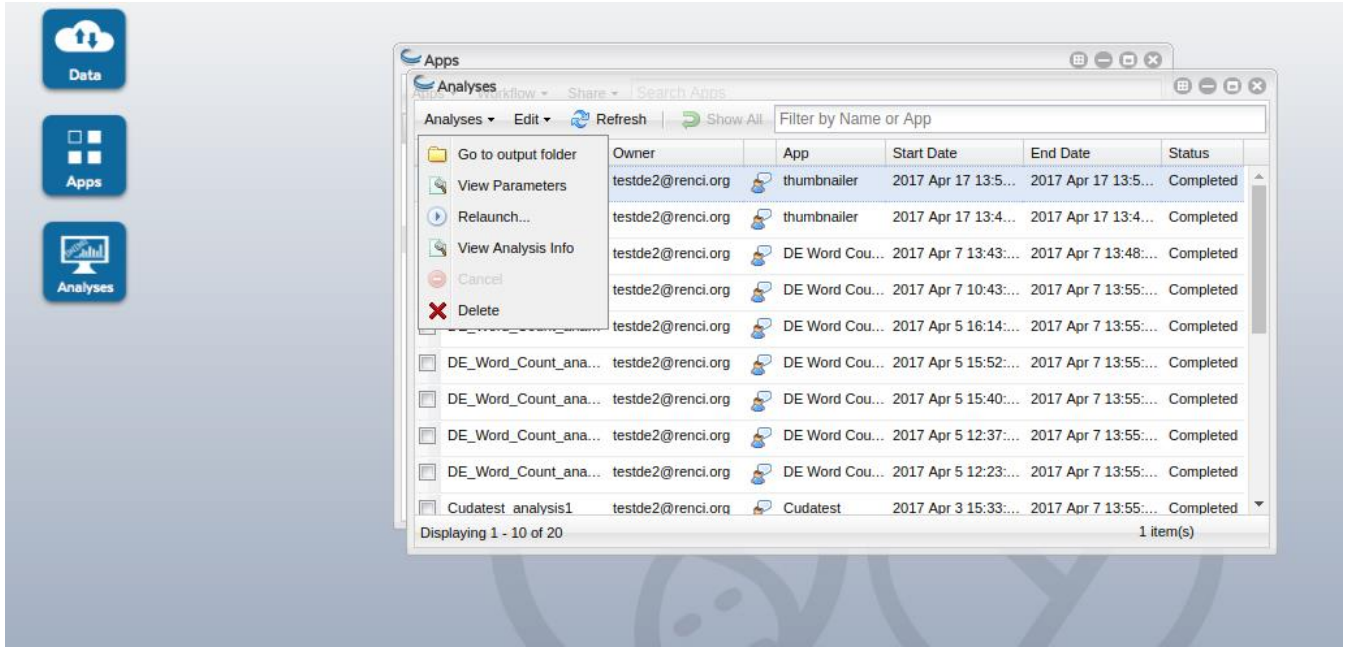
Analysis products, provenance metadata, parameters appear in the grid when complete

---

640 cores → 2.5 hours

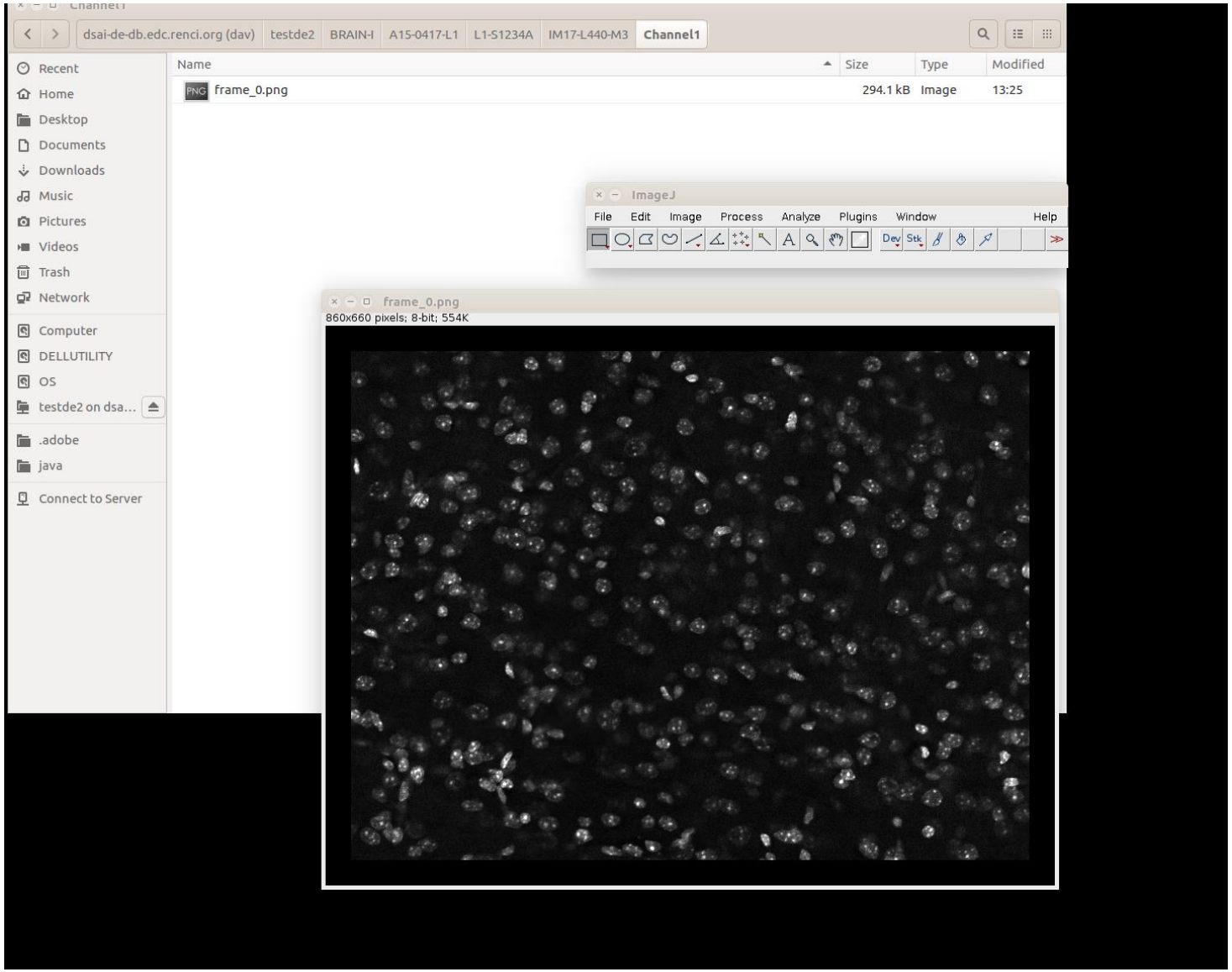
3072 cores → 38 minutes

GPU clusters - - - - → <5 minutes

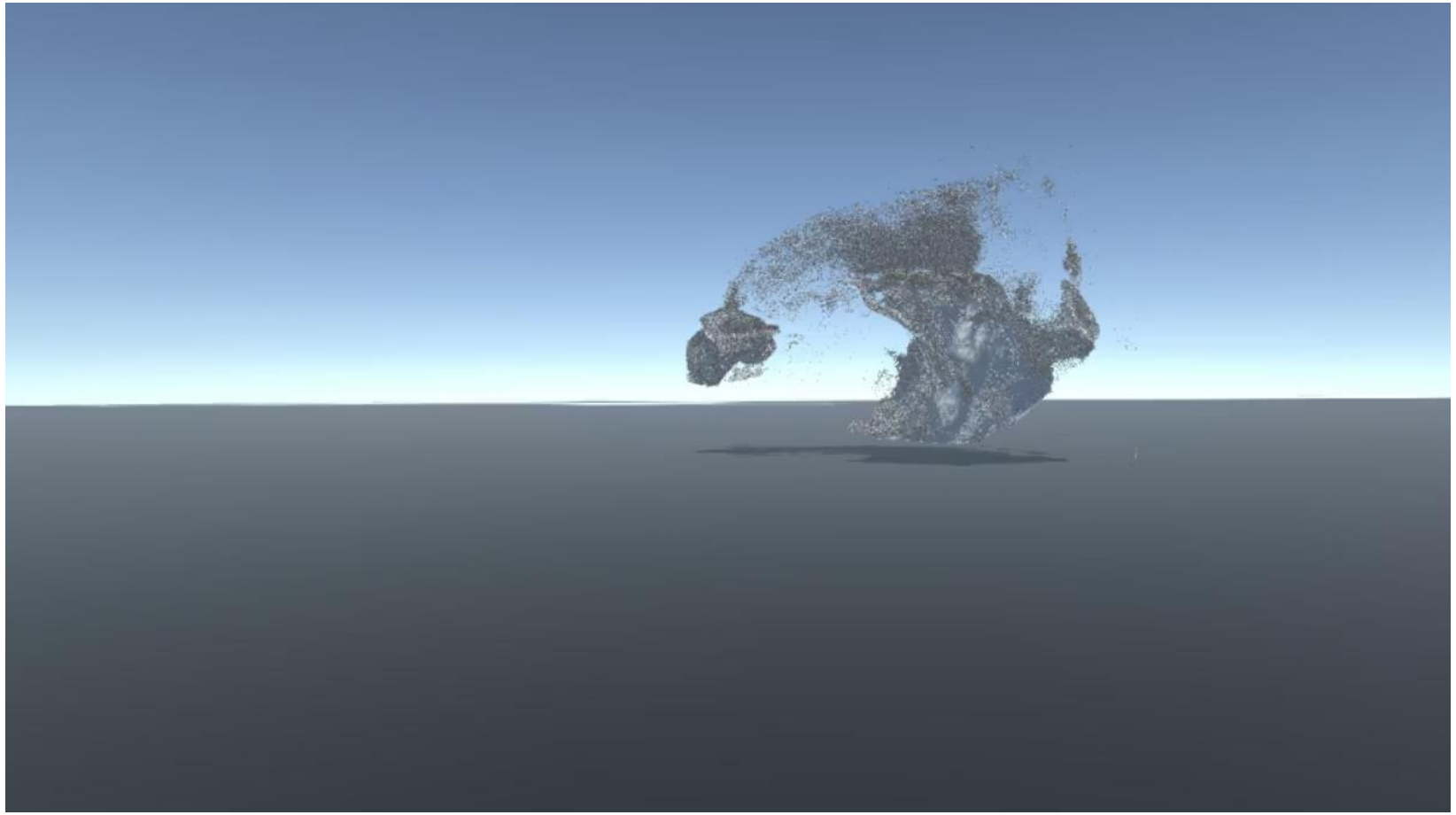


Easy desktop/web access for researchers

- Data grid integrates with desktops and common domain tools.
- Here we are viewing BRAIN-I data on a desktop using off-the-shelf image tools such as ImageJ
- Plan to add access via Jupyter notebooks very soon

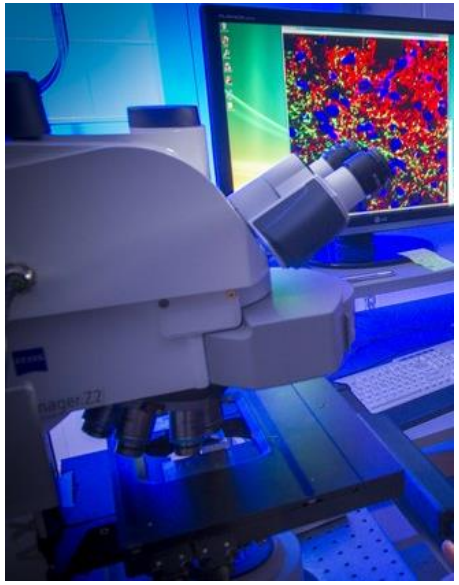


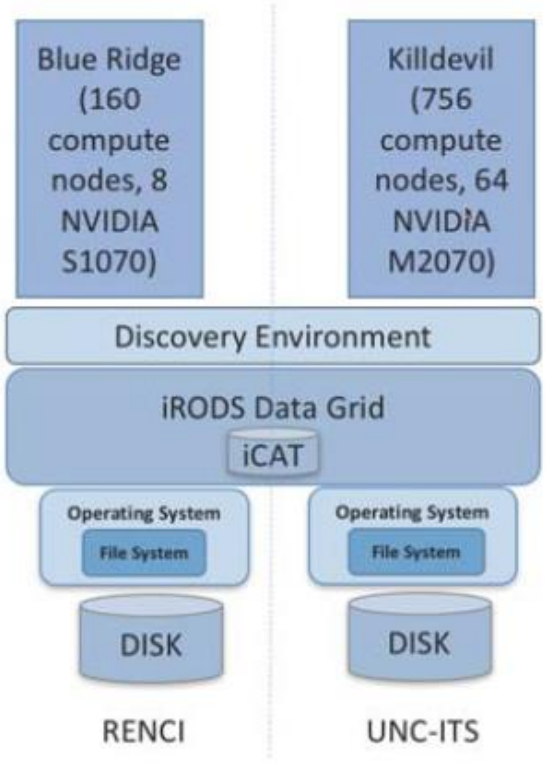
Using Oculus for 3-D Visualization

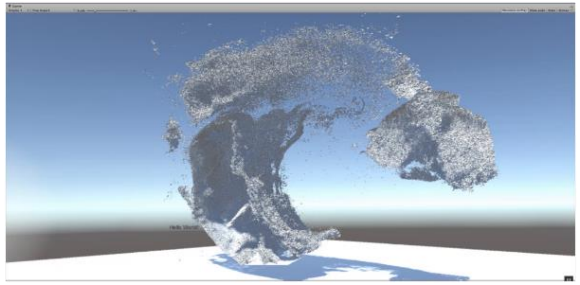


iRODS helps BRAIN-I gets cyberinfrastructure out of the way of science

- Easy, reliable data management and tracking from microscope to publication
- Intuitive environment for computation and data sharing
- Policy based data management, secure and auditable








Surgical Critical Care Initiative (SC2i)
#93587518

SC2i: Surgical Critical Care Initiative
Precision Medicine for Acute Care

- Goal of SC2i: To create clinical decision support tools that focus on best choices for each patient based on data collected from studies at civilian and military research hospitals.
- Partners:
  - Uniformed Services University of the Health Sciences
  - Walter Reed National Military Medical Center
  - Naval Medical Research Center
  - Duke University School of Medicine
    - RENCI is a sub-contractor to Duke
  - Emory University School of Medicine
  - Decision Q
  - Henry M Jackson Foundation for the Advancement of Military Medicine

See: www.sc2i.org

Central Data Repository (CDR) in SC2i

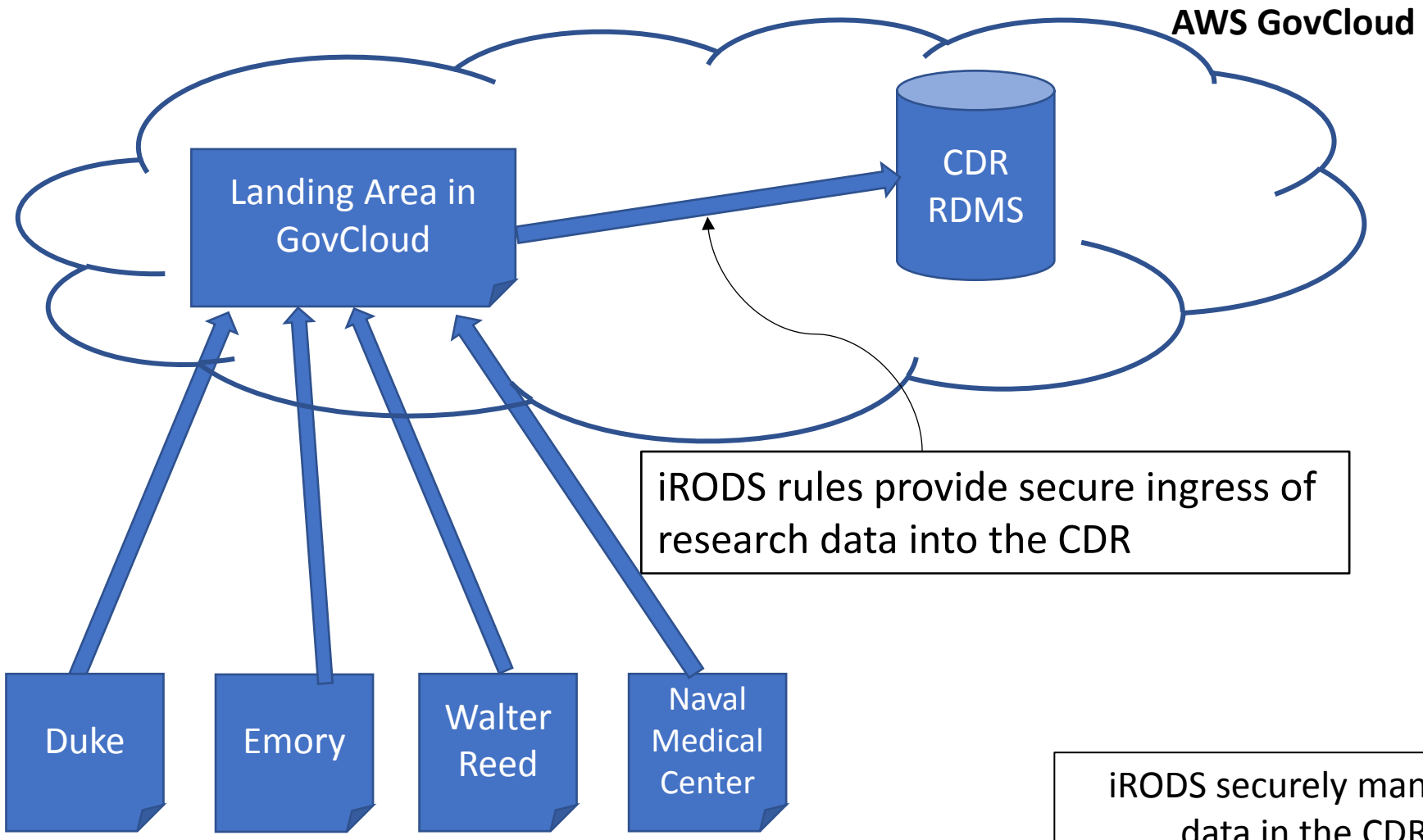
- Data from all institutions is saved in a Central Data Repository for analysis and visualization.
- RENCI is primarily responsible for architecting, implementing and maintaining the CDR
- The CDR is a secure system in AWS GovCloud

- GovCloud is a FedRAMP compliant region within Amazon Web Services (AWS)
- Provides secure/compliant infrastructure for government customers
- CDR runs on GovCloud infrastructure

FedRAMP
Federal Risk Authorization Management Program

FedRAMP is government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services

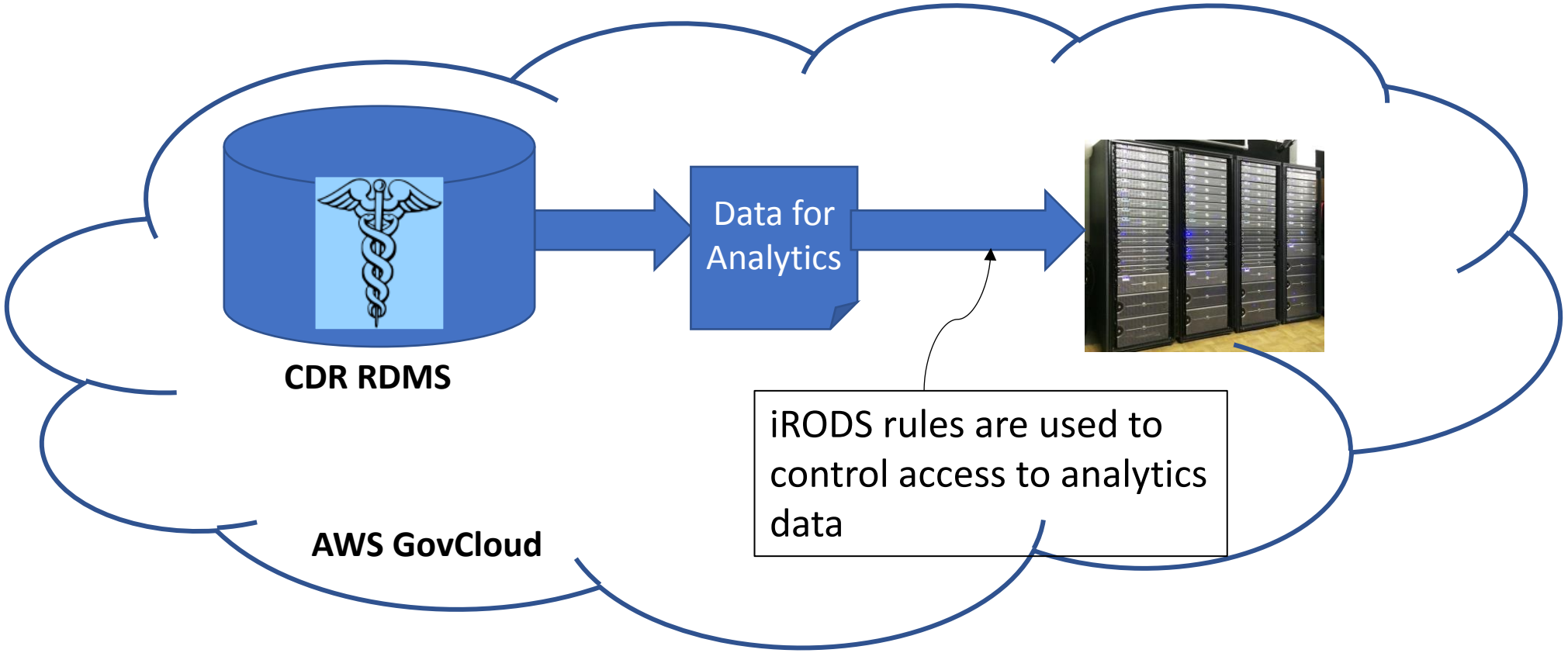
Data Upload and Ingest Using iRODS



iRODS's configurable access control, customizable rules and policies, and secure user management features fulfill security and privacy requirements

iRODS securely manages data in the CDR

Data ETL for Analytics using iRODS



Contact

Ashok Krishnamurthy
Deputy Director
RENCI

ashok@renci.org