Globus Automate

Ian Foster, Ryan Chard, Kyle Chard
Argonne National Laboratory & University of Chicago
foster@anl.gov, rchard@anl.gov, chard@uchicago.edu

globus labs

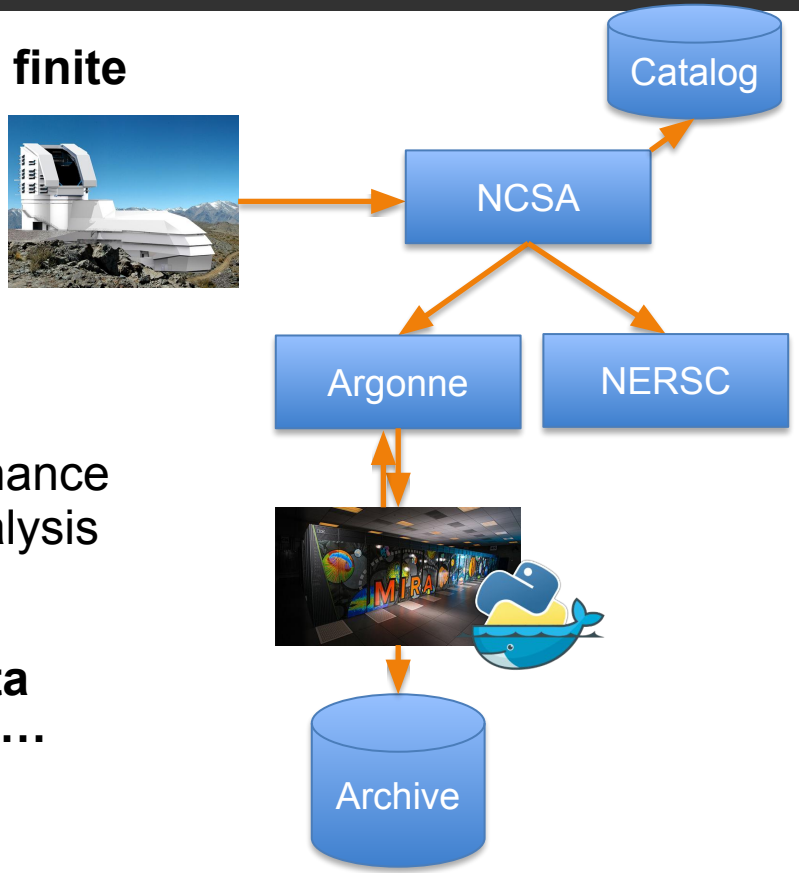
Data management challenges as volumes increase

Data volumes and velocities are overwhelming finite human capabilities

Scientific results are dependent on

- Data acquired at various locations/times
- Analysis processes executed on distributed resources
- Catalogs of descriptive metadata and provenance
- Dynamic collaborations around data and analysis

Best practices are often overlooked, useful data forgotten, errors propagate through pipelines, …



LSST data distribution and analysis pipeline

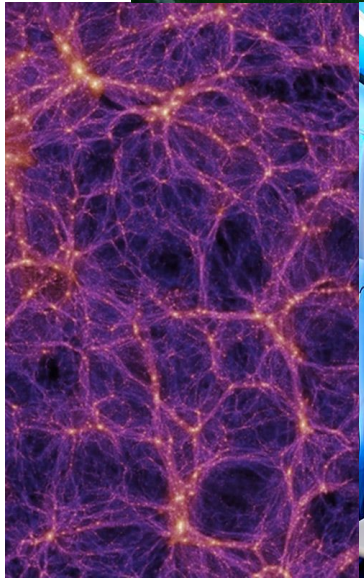
Experimental Science

Data management issues are particularly evident in large scale experimental science

Researchers are allocated short periods of instrument time

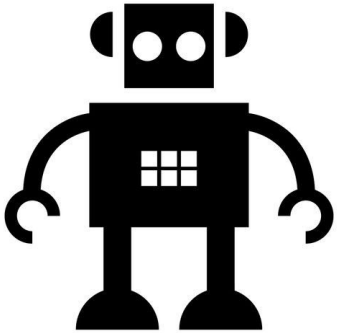
- Must maximize experiment efficiency and output data quality/accuracy

Inefficiencies mean less science is performed and researchers may have to wait months for another chance.



Automation

Goal: Automate data manipulation tasks from transfer and sharing to acquisition, publication, indexing, analysis, and inference



Requirements: A platform that...

- Can automate best practices (replicate, catalog, share)
- Is data driven -- responds as data are created
- Can be applied across arbitrary storage and compute infrastru
- Can be dynamically programmed to respond to new events
- Enable non-expert users to define automations

Approach: Compose and execute data manipulation flows through the Automate PaaS

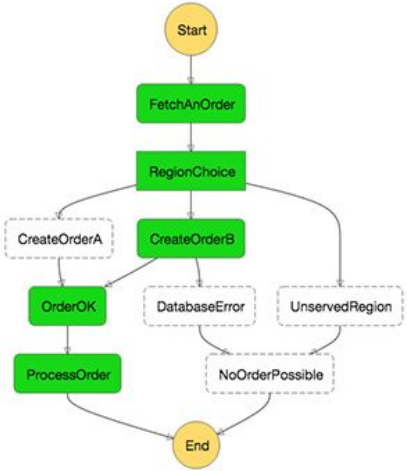
Globus Automate

Built on AWS Step Functions

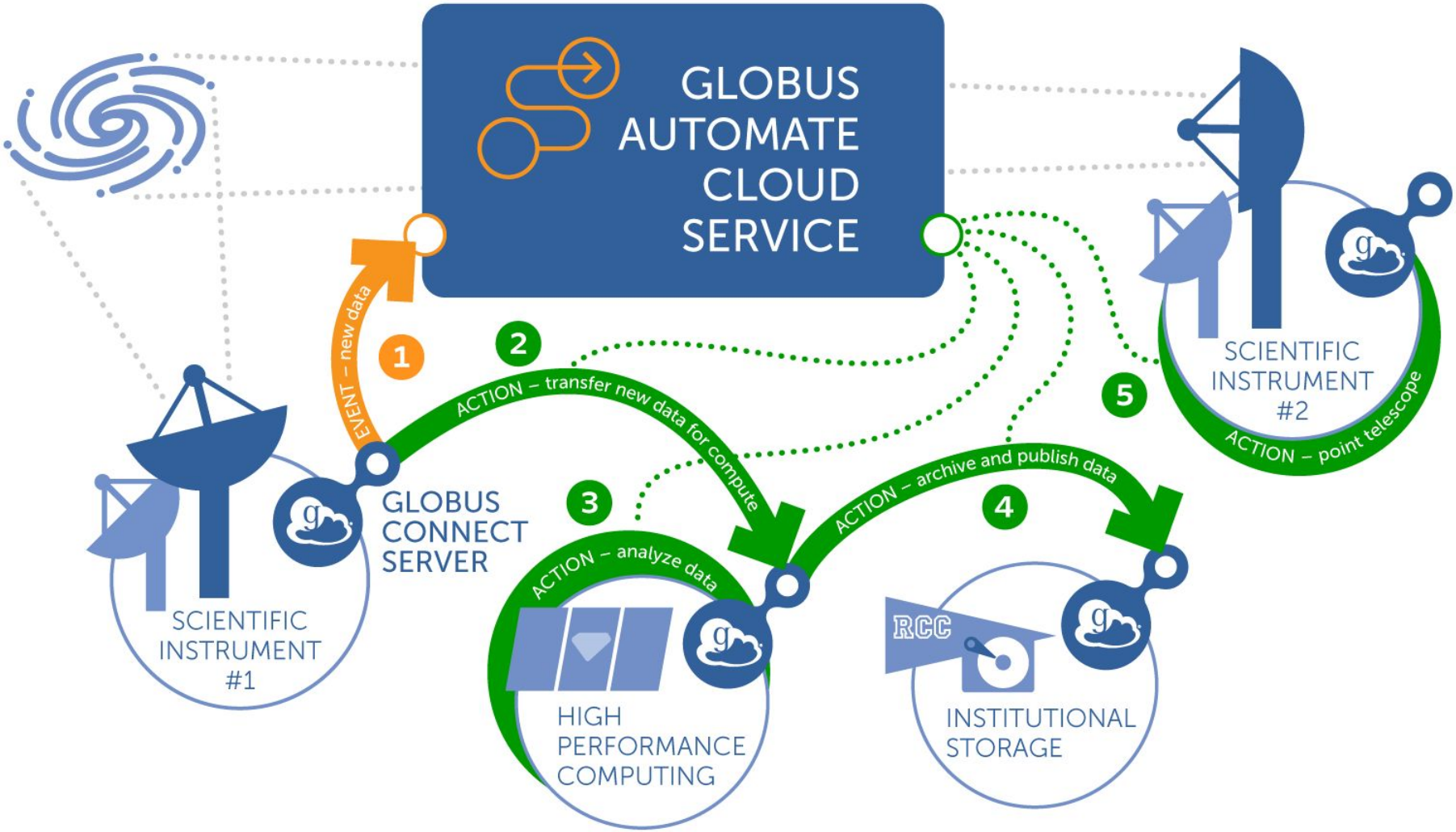
- Simple JSON-based state machine language
- Facilitates conditions, loops, fault tolerance, etc.
- Propagates state through the flow

Standardized API to integrate custom event & action services

- Actions can be either synchronous or asynchronous
- Custom Web forms can halt flows for user input



Actions are secured with Globus Auth


GLOBUS AUTOMATE CLOUD SERVICE
EVENT – new data
1
2
ACTION – transfer new data for compute
3
ACTION – analyze data
4
ACTION – archive and publish data
5
ACTION – point telescope
GLOBUS CONNECT SERVER
SCIENTIFIC INSTRUMENT #1
HIGH PERFORMANCE COMPUTING
RCC
INSTITUTIONAL STORAGE
SCIENTIFIC INSTRUMENT #2

Automate Prototype: The Service

Define JSON flows that step between action services and describe JSON doc of default input data

- Definition based on AWS state machine language

Associate a trigger condition -- event data is passed in when executed

We provide a polling SFN activity you can use to halt a flow until an action_id has completed

Automate Prototype: Actions

Any service can expose the Action API

- /automate/v1/action/run, status, cancel, introspect, …
- .../status used to enable polling
- We give the service an action_id on invocation

When registering an action we make an internal lambda function that calls your service's url

- Makes an ARN for it and maps to a user-friendly name for use in flows

Introspect tells us what input the action accepts -- used during flow creation

The action can then be stepped to in a flow

Some Actions

Auth

- User login
- Secure service interactions
- App identity and interactions

Identifier

- Manage namespace
- Mint DOI

Search

- Catalog
- Faceted search
- Search query

Transfer

- File operations
- Transfer data
- Set permission

Execute

- Remote execution
- Secure connections
- Self optimization

Automate Prototype: Events

Any service can expose the Events API

- /automate/v1/event/register, poll, introspect, …

Automate polls each event interface and adds responses to a reliable Simple Queue Service queue

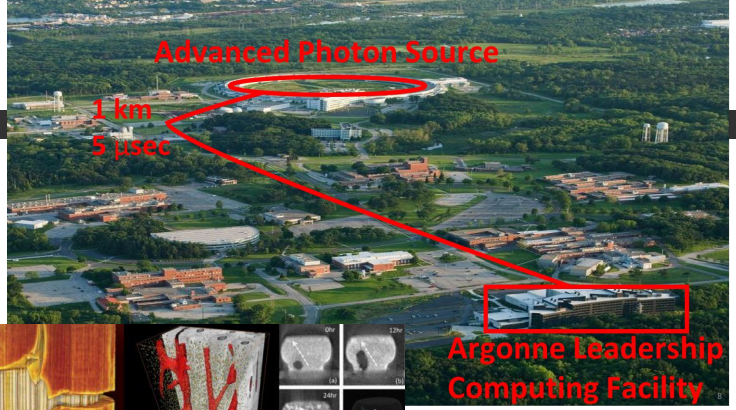
- Events processed by lambda functions

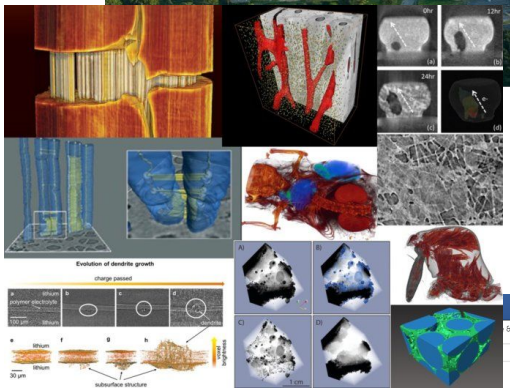
Integrates Ripple/Dash as an event source

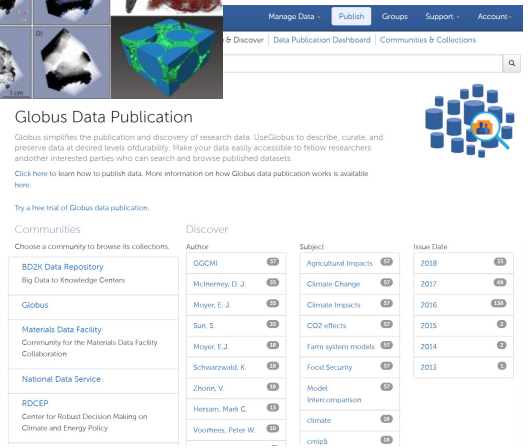
- Can be driven by file events

Use Cases

- Neuroanatomy (APS)
- Advanced Light Source
- Data Publication



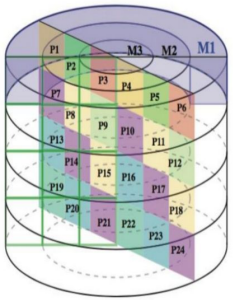


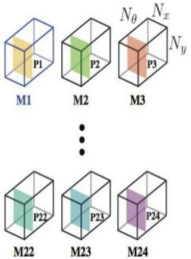


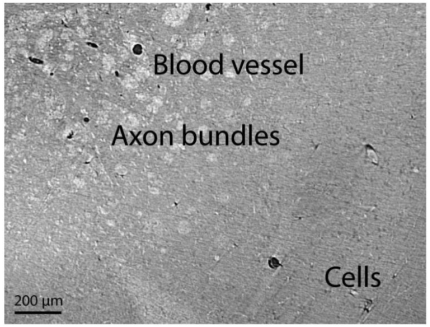
Case 1: Neuroanatomy

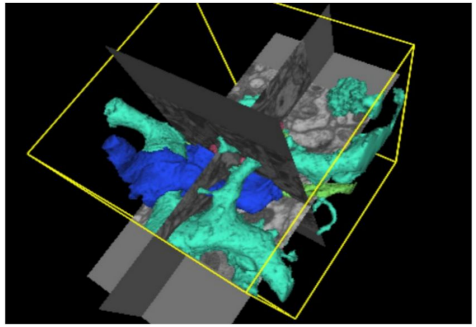
UChicago's Kasthuri Lab study brain aging and disease

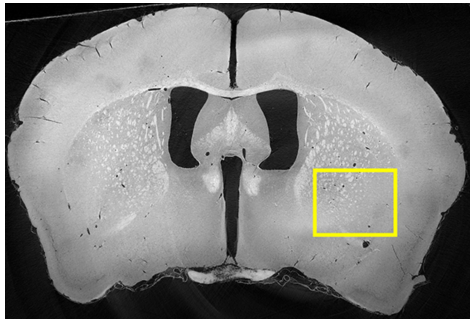
- Construct connectomes -- mapping of neuron connections
- Use synchrotron (APS) to rapidly image brains (and other things)
- Given beam time once every few months
- Generate segmented datasets/visualizations for the community
- ~20GB/minute for large (cm) unsectioned brains
- Perform semi-standard reconstruction on all data across HPC resources



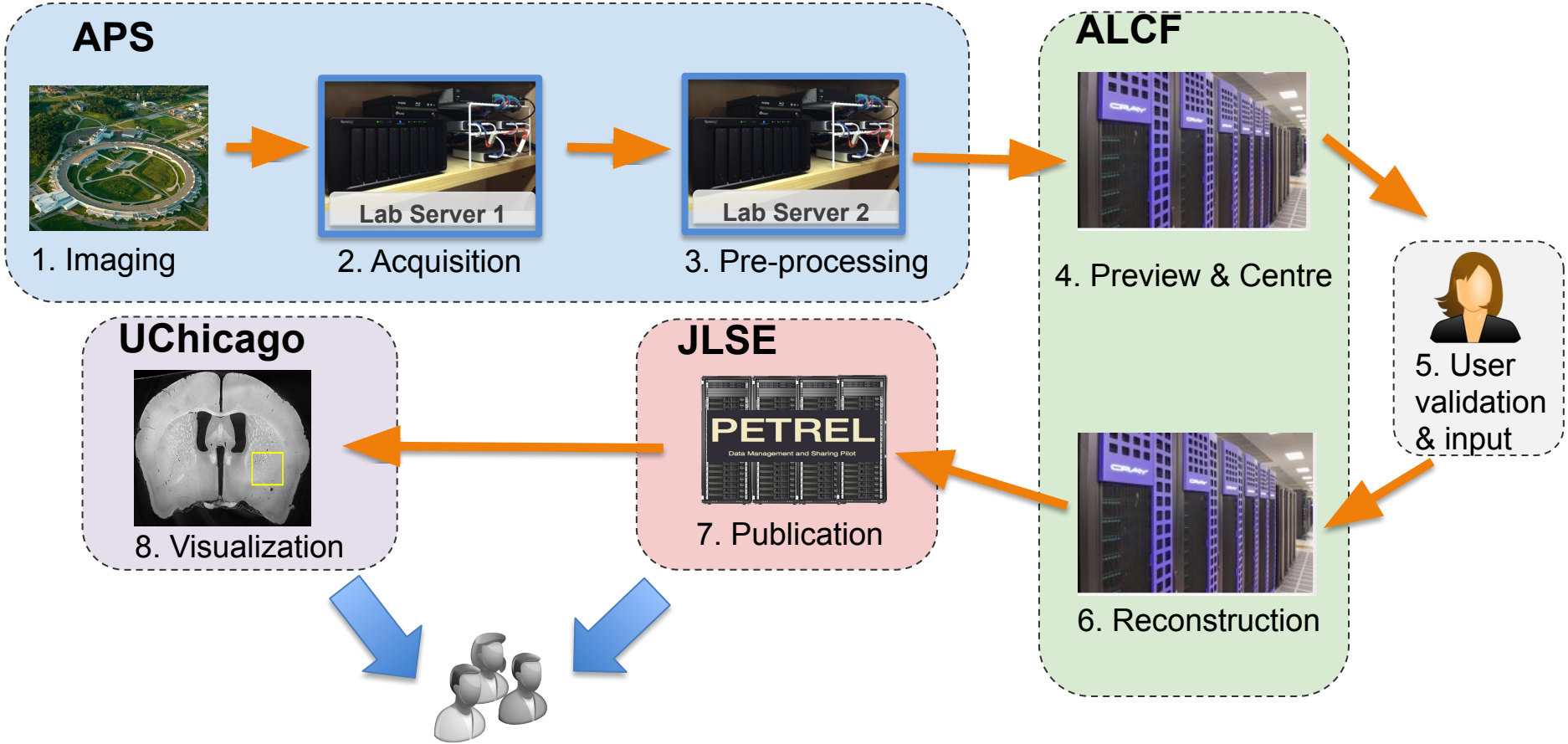




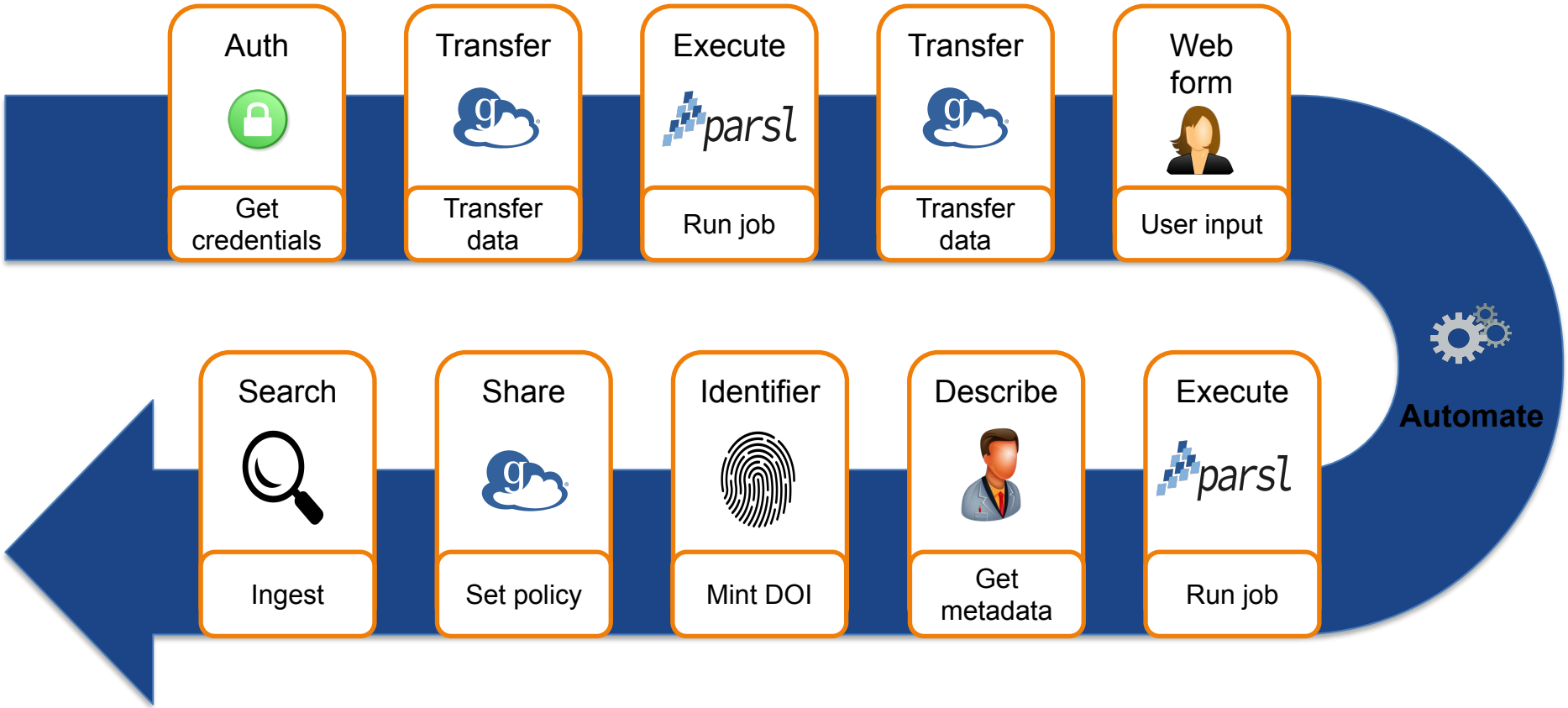




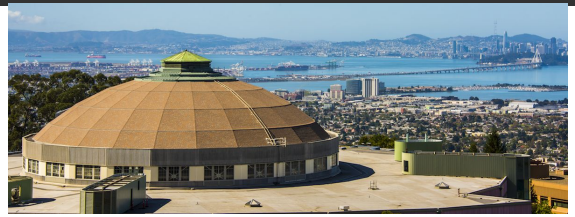
Neuroanatomy Overview

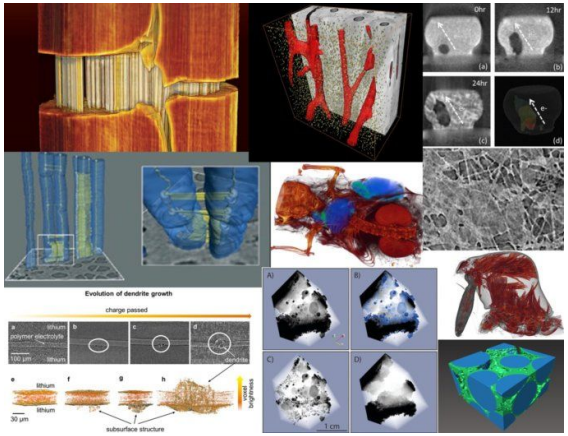


Neuroanatomy Automation



Case 2: Advanced Light Source





Reconstructions

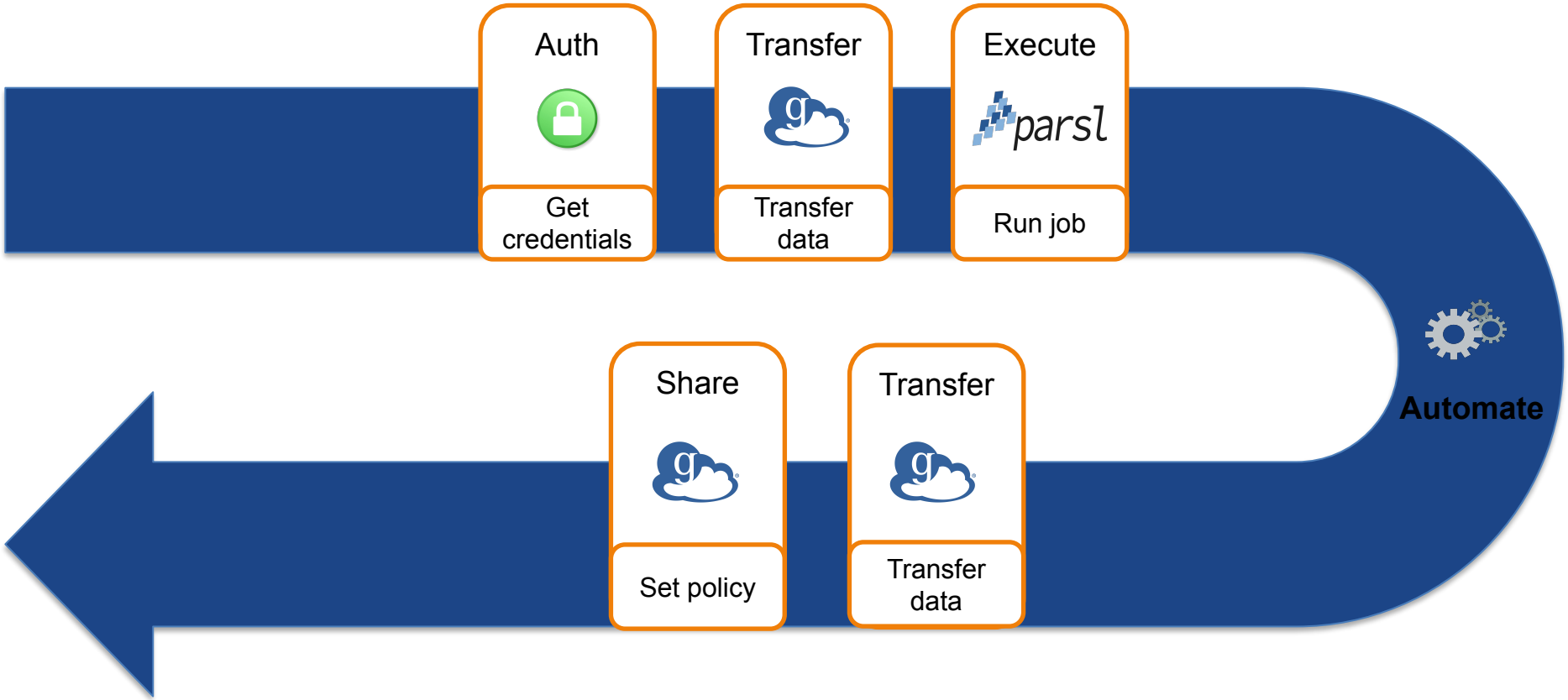
- Move data to NERSC
- Submit batch reconstruction
- Return result to users

Requirements

- Plug in different tool, endpoint, allocation

Leverage multiple compute resources (NERSC, local, AWS)

ALS Automation



Case 3: Data Publication

Citable Data

Standard metadata,
persistent identifiers,
durable storage

Institutional Data

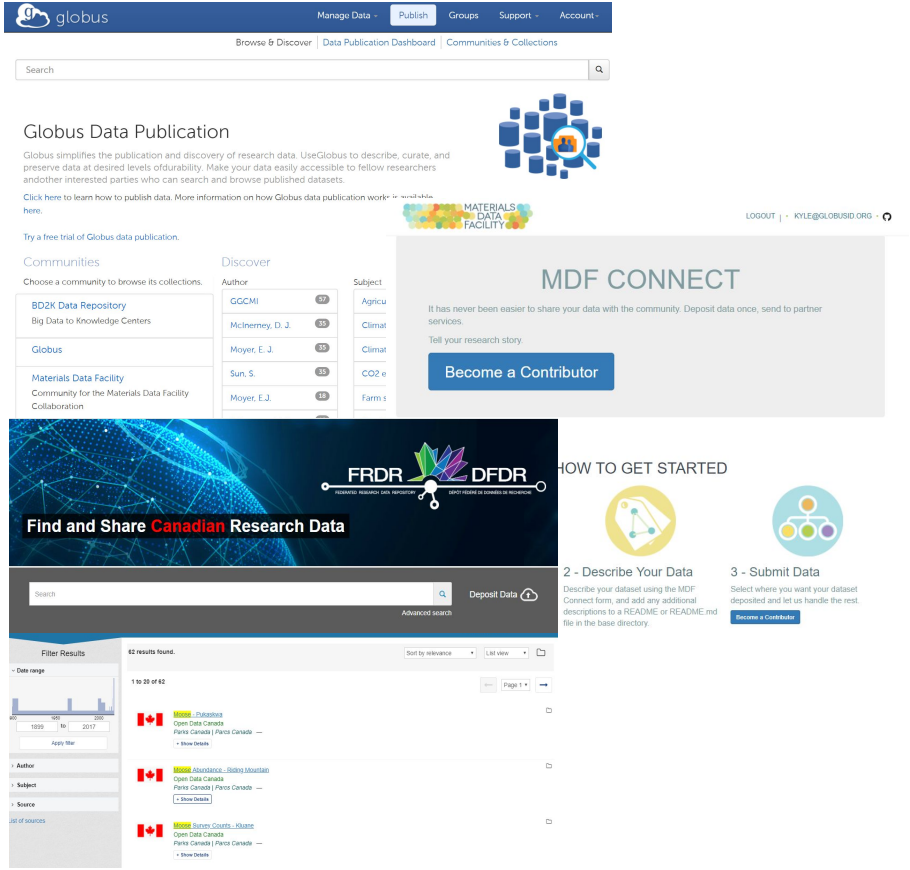
Many domains,
custom metadata,
locally managed storage

Community Data

Agreed schema,
larger datasets,
fine grained metadata

Globus Publication v1

- Cloud-based web app
- BYO storage & in-place publication
- User-managed collections
- Select pre-defined schema
- Handle, DOI persistent identifiers
- Adoption since 2015:
  - >2000 users, >600 datasets



Publication v2 via Automate

- Decompose Globus Publish v1 into platform services
- Allow for flexible re-composition and adaptation pf services
- Enable extension and enhancement