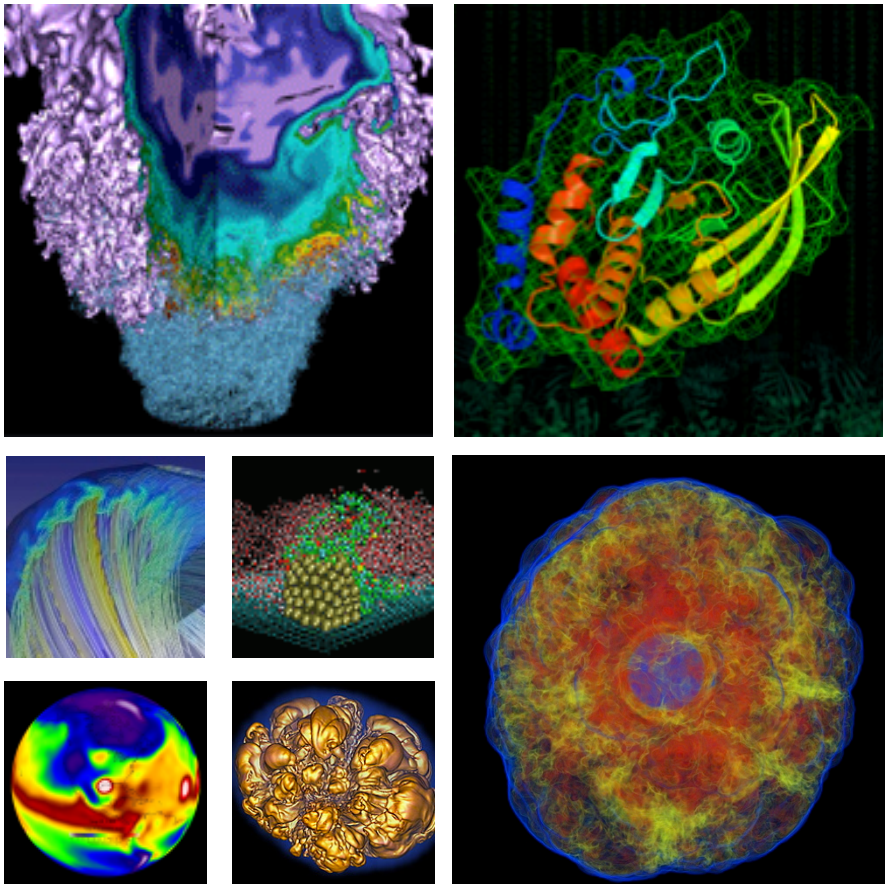
Integrated tools for NGBI



NERSC

Joaquin Correa, OSP

September 4, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC :
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING

Computing and Data for science

- 5000 users, 600 projects
- From 48 states; 65% from universities
- Hundreds of users each day
- ***1600 publications per year***

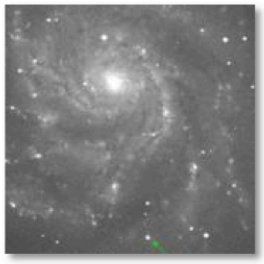
Systems designed for science

- 1.3PF Petaflop Cray system, Hopper
  - 8th Fastest computer in US
  - Fastest open Cray XE6 system

RECENT NERSC User Scientific Accomplishments

Astrophysics

NERSC played a key role in the discovery that led to the 2011 Nobel Prize in Physics.
(S. Perlmutter, UC Berkeley/LBNL)

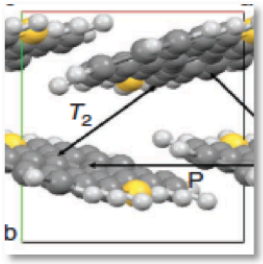


Astrophysics

The earliest-ever detection of a supernova was made possible by NERSC and ESnet.
(P. Nugent, LBNL)

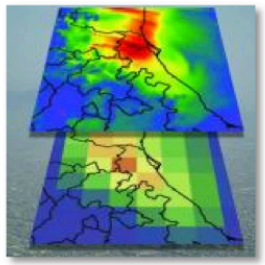
Materials

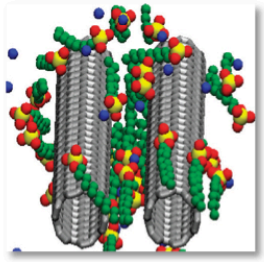
A vastly improved organic semiconductor discovery is a key proof of principle for rational design of new materials.
(A. Aspuru-Guzik, Harvard)



Climate

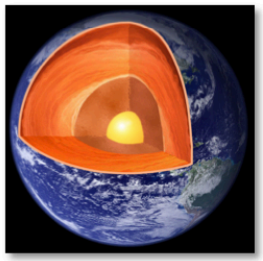
Atmospheric scientists have shown how small-scale effects of aerosols contribute to errors in climate models.
(W. Gustafson, PNNL)





Chemistry

Molecular dynamics simulations show how certain surfactants can be used to separate out bundles of carbon nanotubes with important



Nuclear Physics

The KamLAND neutrino experiment showed that radioactivity cannot be Earth's only heat source; it accounts for only ½ of it.

Integrated tools for NGBI

- **What does NERSC's NGBI team do?**

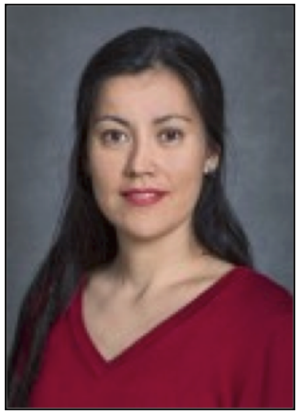
We focus on enabling high-performance computational solutions adapted to bio-imaging applications through an integrated highly accessible **shared scalable web-service that provides a one-stop-shop for producers and consumers of models built on imaging data** to refine pixel data into actionable knowledge resources.

- **Who is involved?**

David Skinner

Shreyas Cholia

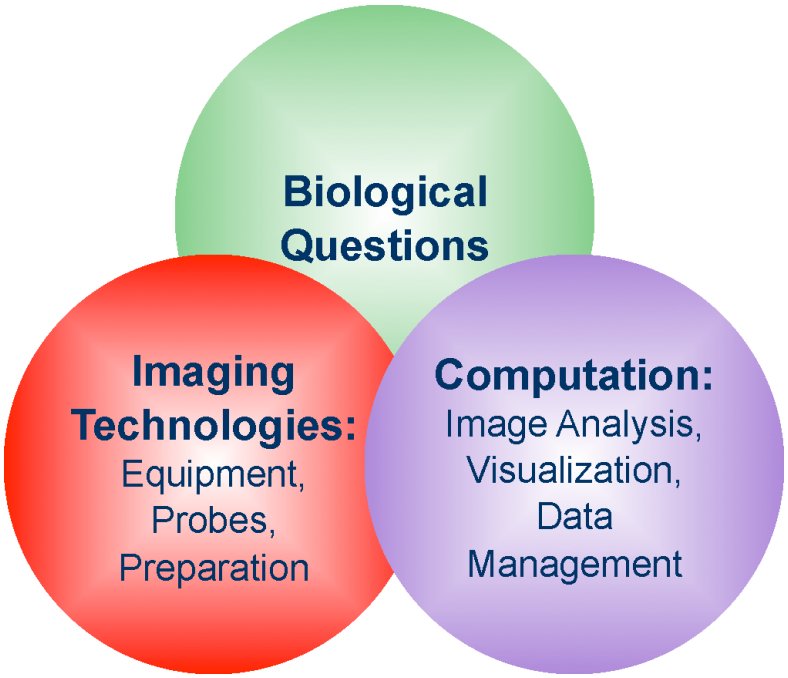
Joaquin Correa



Manfred Auer *Damir Sudar* *Dani Ushizima* *Joerg Meyer*

NERSC	LSD	CRD

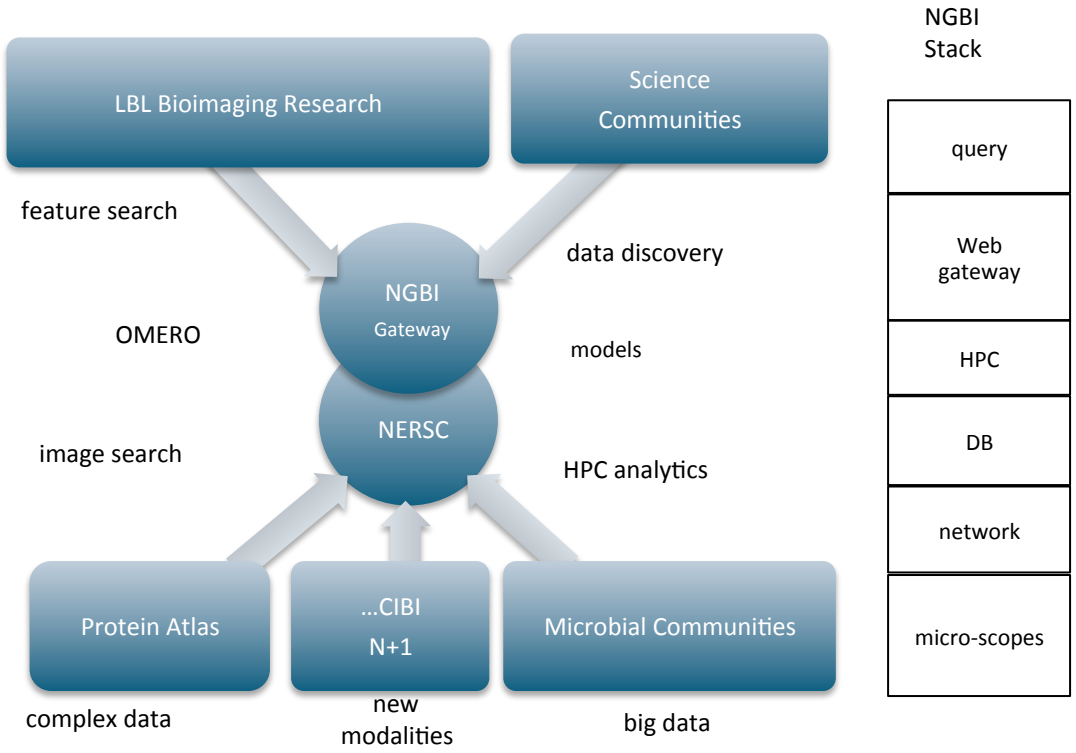
Key scientific issue being addressed



- Building biological models from massive streams of pixel data
- Spatial understanding of protein expression (**PA**)
- Statistical understanding of microbial communities in biofilms (**MC**)
- Improving imaging modalities from emerging instruments

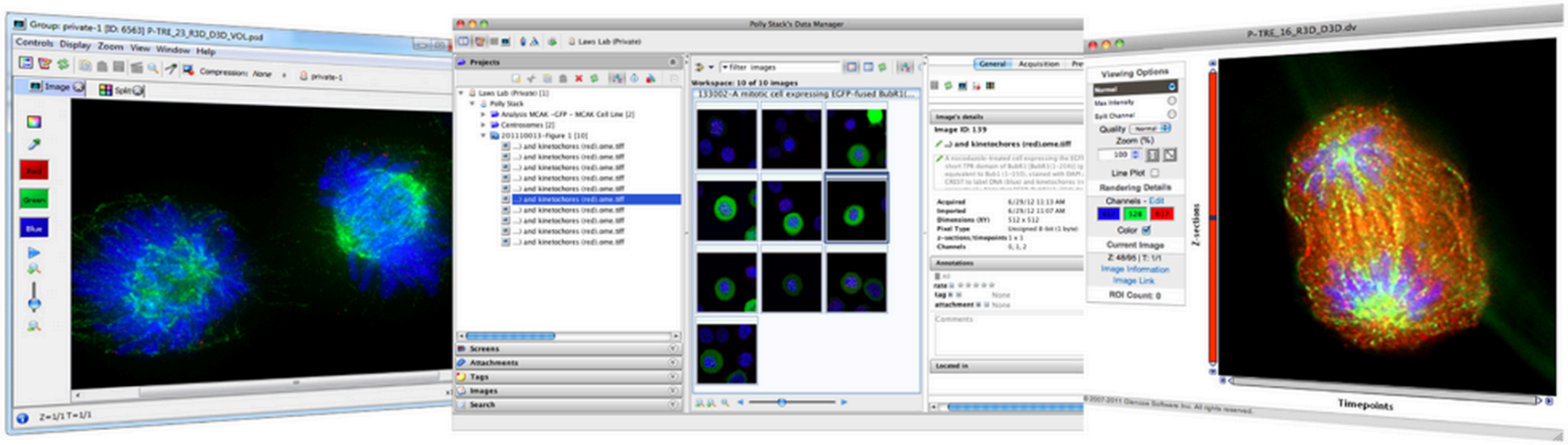
Approach

- One-stop-shop for building and using biological models built on image data
- Leverage scalable computing and data techniques, end-to-end data management
- Bootstrapping ML-based image processing with expert crowds. Automating segmentation and classification.
- Co-design of instrumentation and computation



Our solution? An OMERO-based platform

… A shared scalable web-service that provides a one-stop-shop for producers and consumers of models built on imaging data to refine pixel data into actionable knowledge resources …



Import

Over 100 file formats supported, including all major microscope formats.

Organize

Organize by 'Project', 'Tags' or acquisition date. Annotate with attachments, comments, ratings.

View

Move through multiple dimensions, copy and apply rendering settings, zoom and pan 'Big' images.

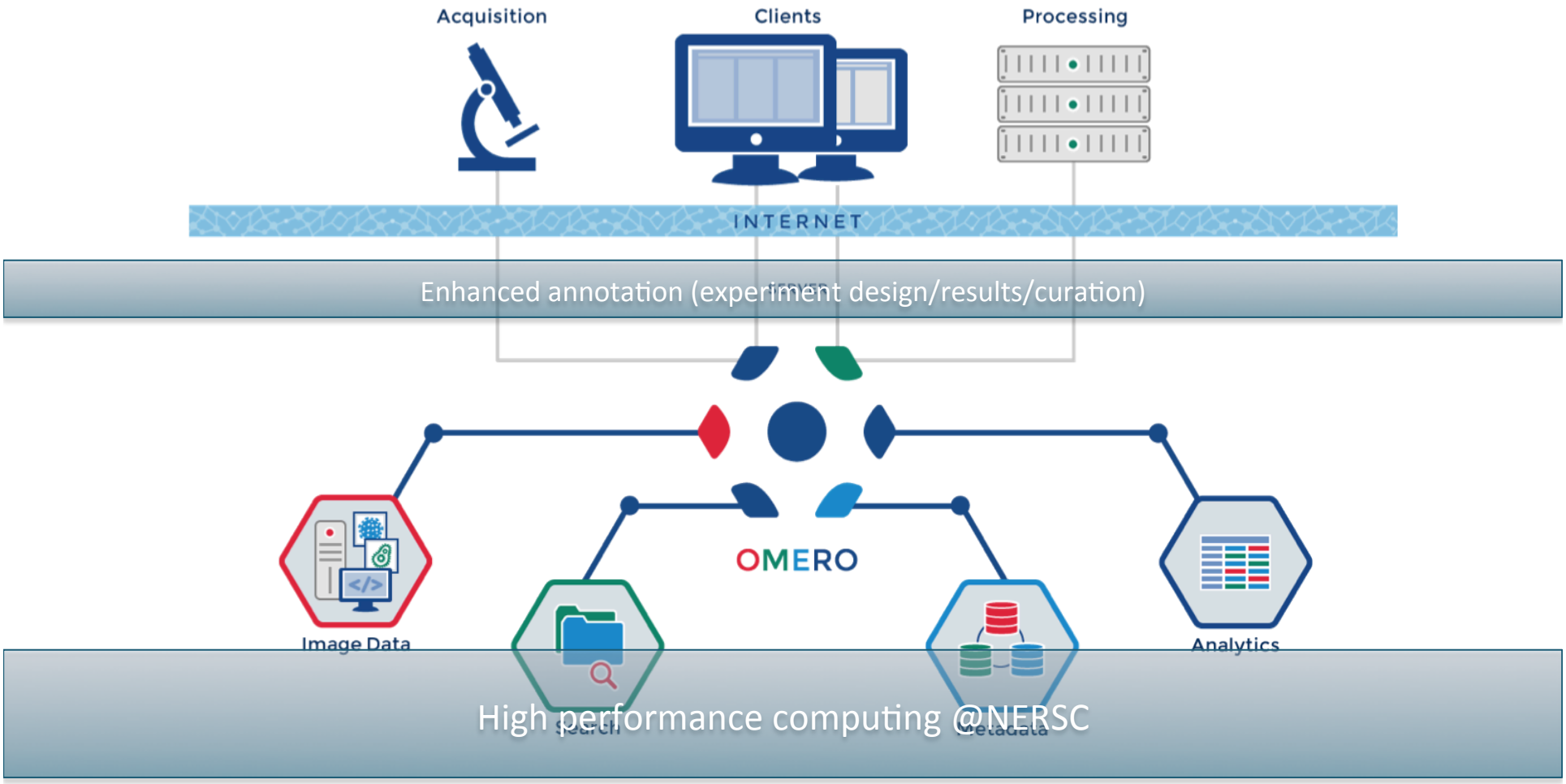
Analyze

Draw and measure regions of interest, write Python scripts in OMERO or connect to OMERO from your favorite analysis software.

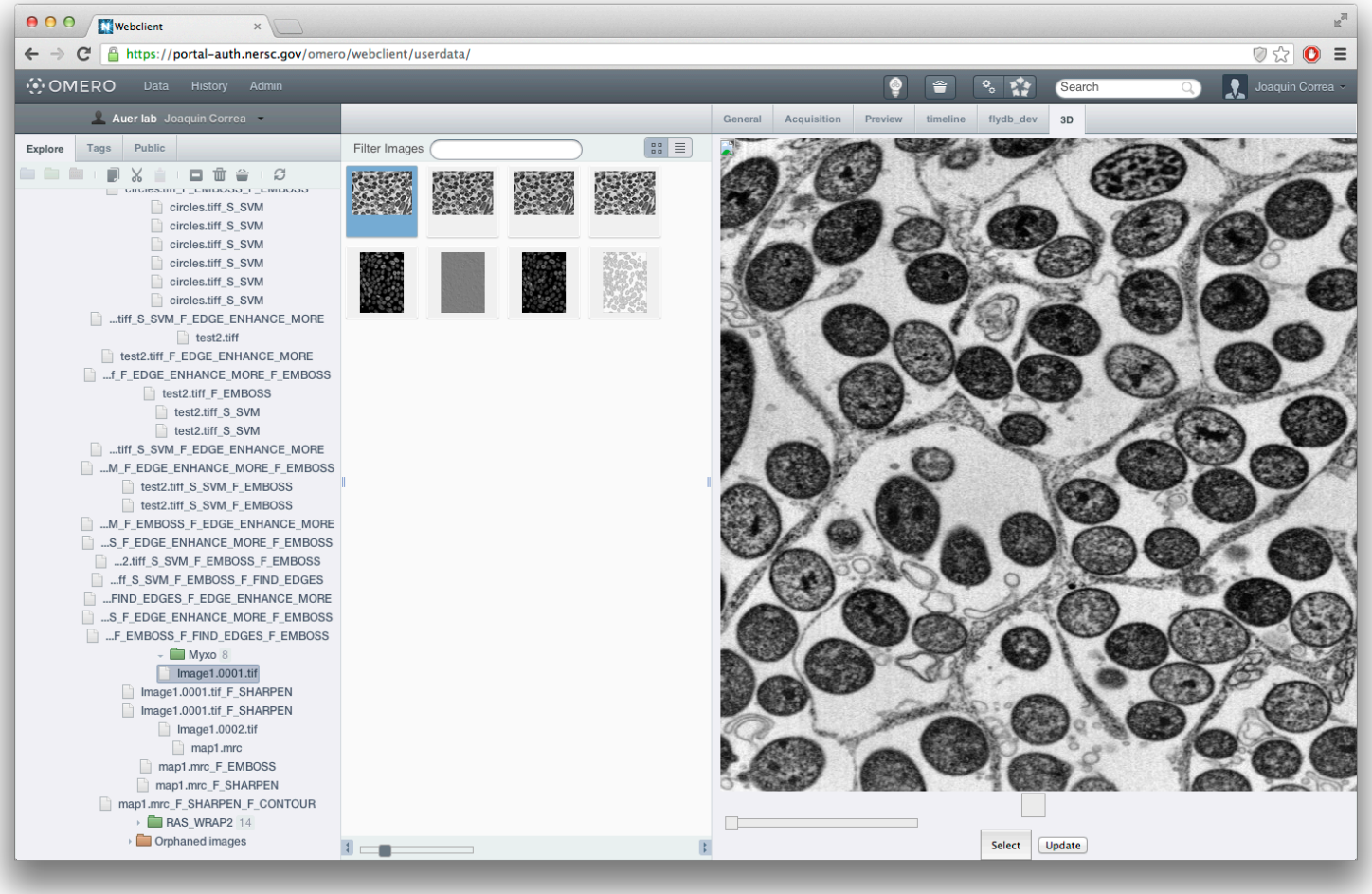
Export

Export your images for analysis or publication. Save images as 'figures' ready for presentations.

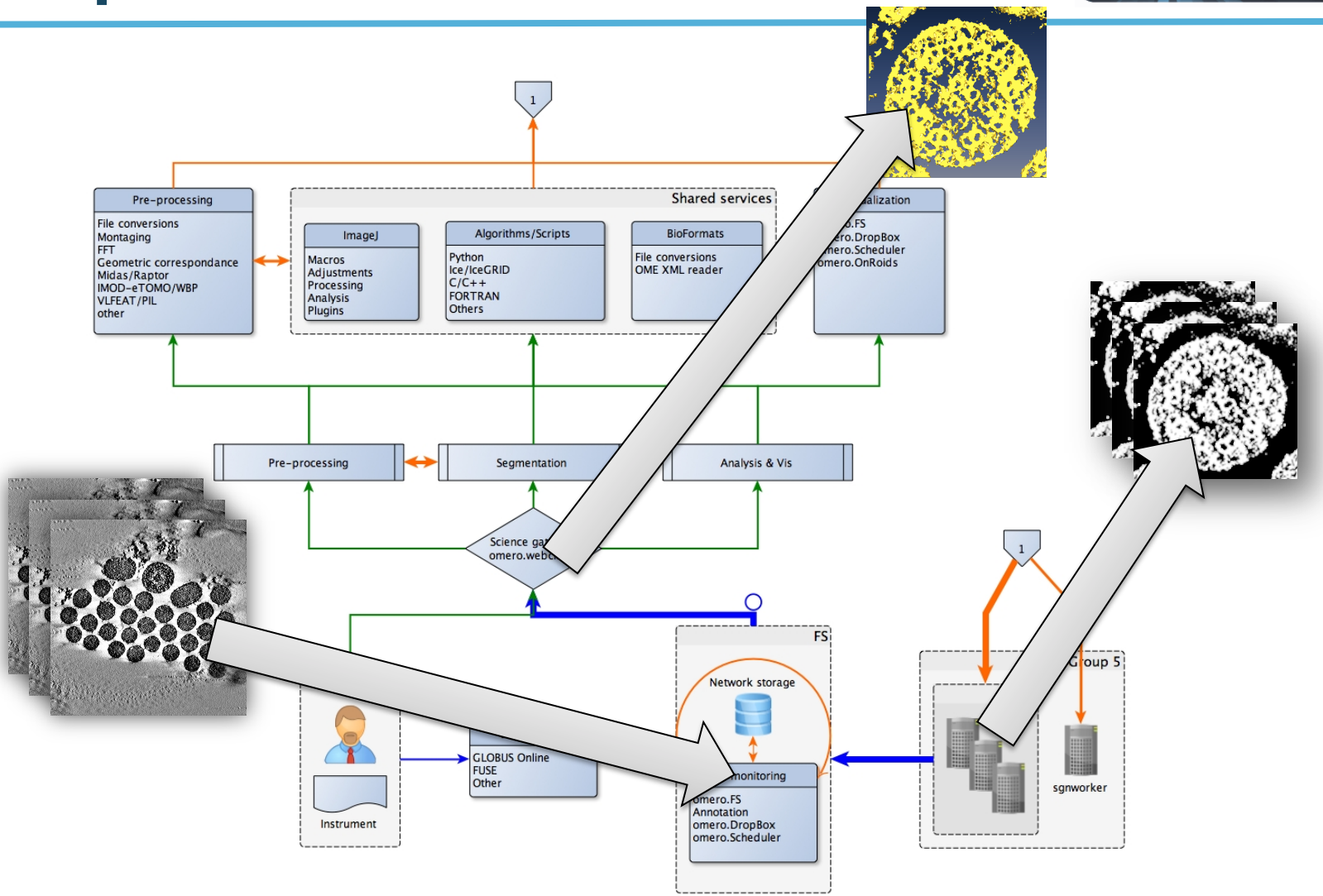
Our solution? An OMERO-based platform



Our solution? An OMERO-based platform



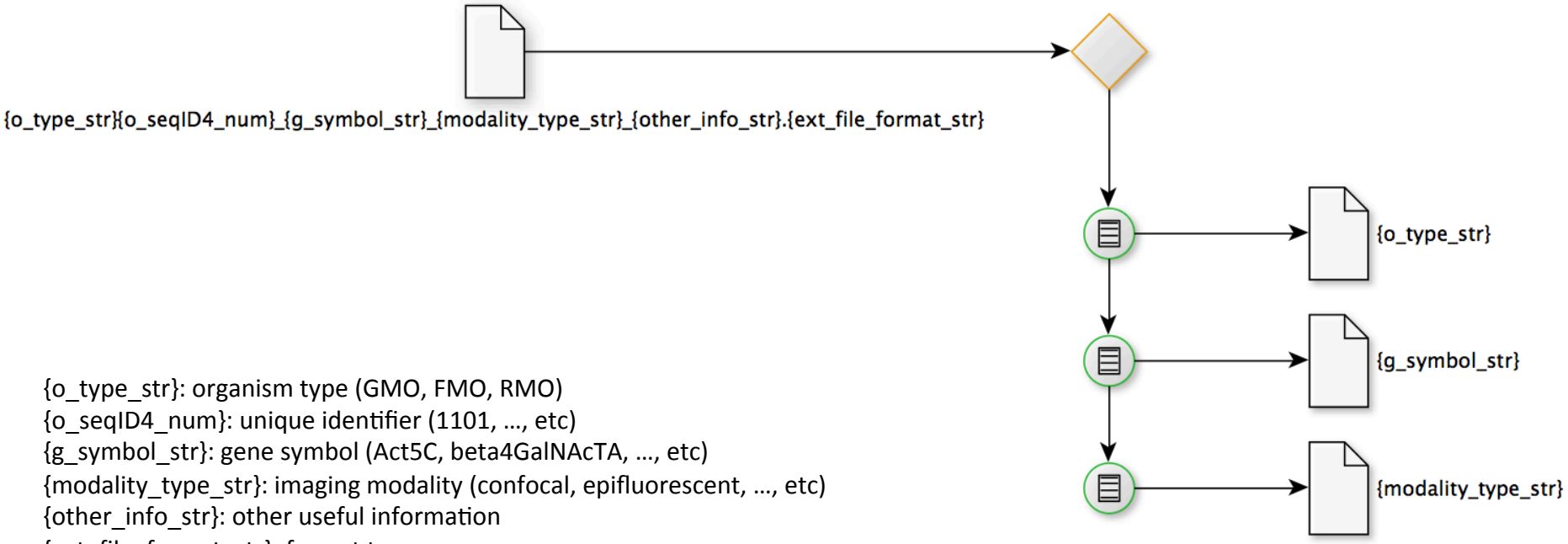
HPC implementation -Architecture



Automated workflows

- Experiment design
- Automated annotation before classification
- Run scripts on the fly or demand-based
- Drag and drop files/folders/archives

Automated annotation, FN-*based*



{o_type_str}: organism type (GMO, FMO, RMO)
{o_seqID4_num}: unique identifier (1101, …, etc)
{g_symbol_str}: gene symbol (Act5C, beta4GalNAcTA, …, etc)
{modality_type_str}: imaging modality (confocal, epifluorescent, …, etc)
{other_info_str}: other useful information
{ext_file_format_str}: format type

EX:

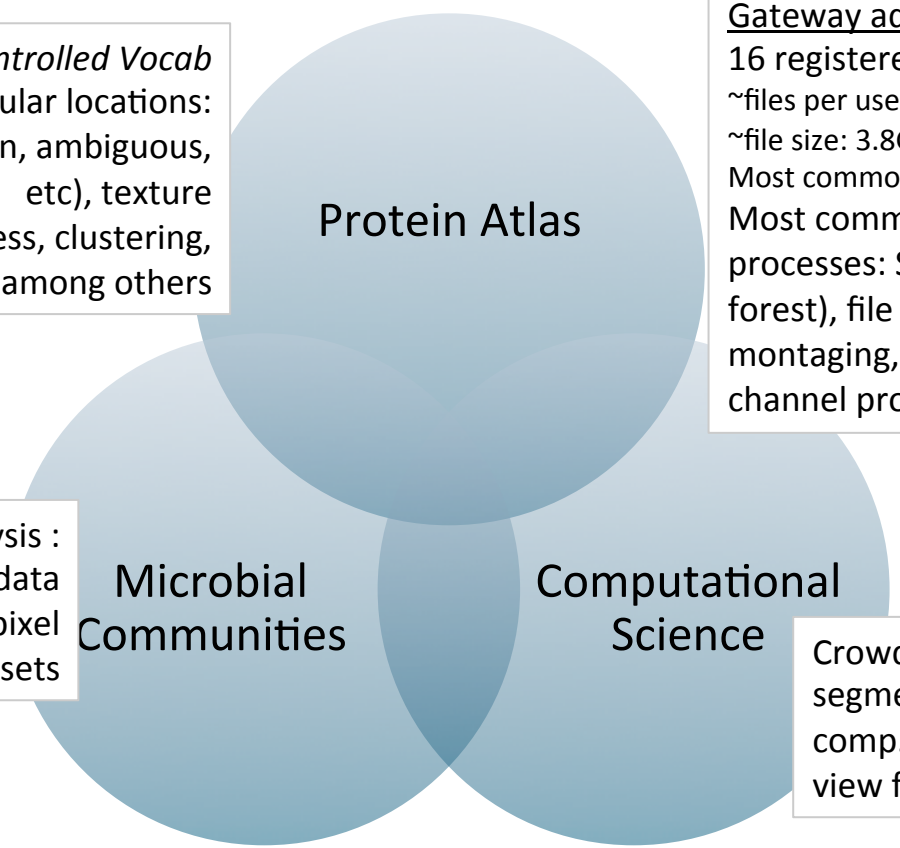
GMO1101_Act5C_CON_updated.lsm

Pixels to knowledge – P2K

High-throughput, high-performance image processing

In 2013 we staffed, developed and deployed custom-tailored image-processing algorithms and successfully implemented a suite of general-purpose processing software using NERSC's high performance computing and its science gateway.

Enhanced annotation - *Controlled Vocab for Model Building (*subcellular locations: centrosome, euchromatin, ambiguous, etc), texture smoothness, roughness, clustering, among others



Gateway adoption

16 registered users from 3 labs

~files per user: 2k-2.5k (stackable)

~file size: 3.8G-5G (per stack)

Most common file type: 3D, 3D+*c* MRC

Most common computing processes: Segmentation (random forest), file conversion, montaging, maximum intensity channel projection

Scalable Image Analysis : segmentation of large EM data sets at NERSC, giga-pixel montaging of data sets

Crowd sourced classification and segmentation to power ML (TBI), comp. engagement, more docs, public view for giga-pixel images

Working on Joint Publication Fall 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

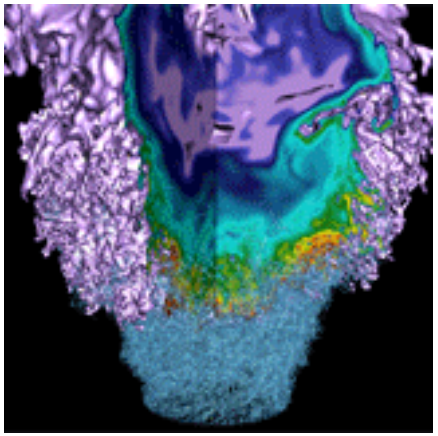
BERKELEY LAB

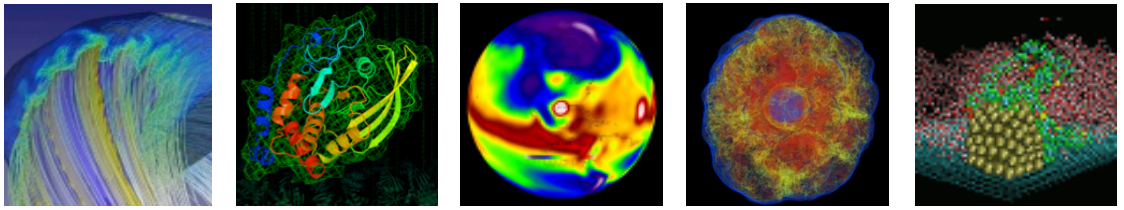
Some numbers

- **No. of registered users:** 16 (3 labs)
- **~files per user:** 2k-2.5k (stackable)
- **~file size:** 3.8G-5G (per stack)
- **Most common file type:** 3D, 3D+*c* MRC
- **Most common processes:** Segmentation (random forest), file conversion, montaging, maximum intensity channel projection

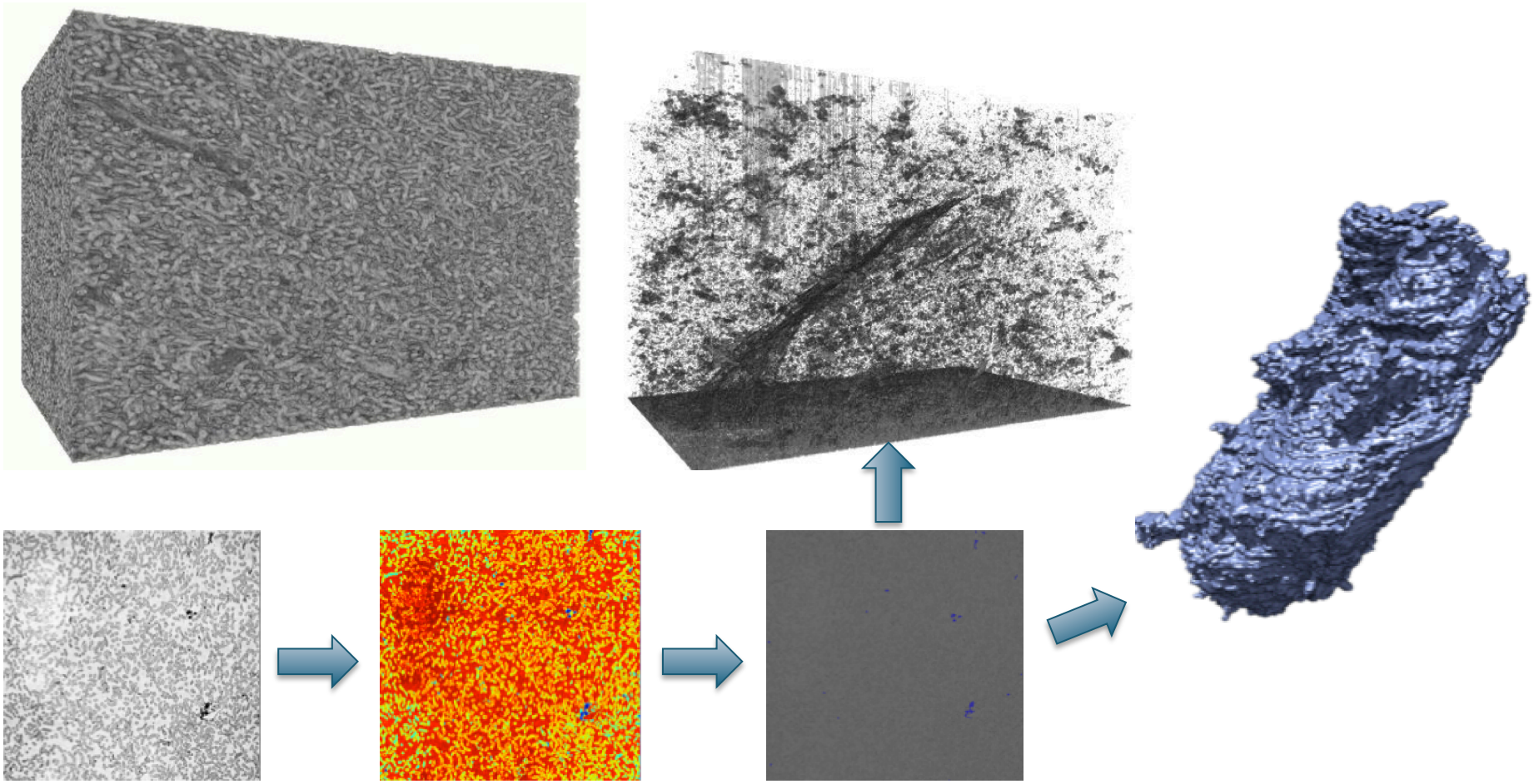
Results

Successful cases: Image pre-processing, segmentation, analysis and visualization





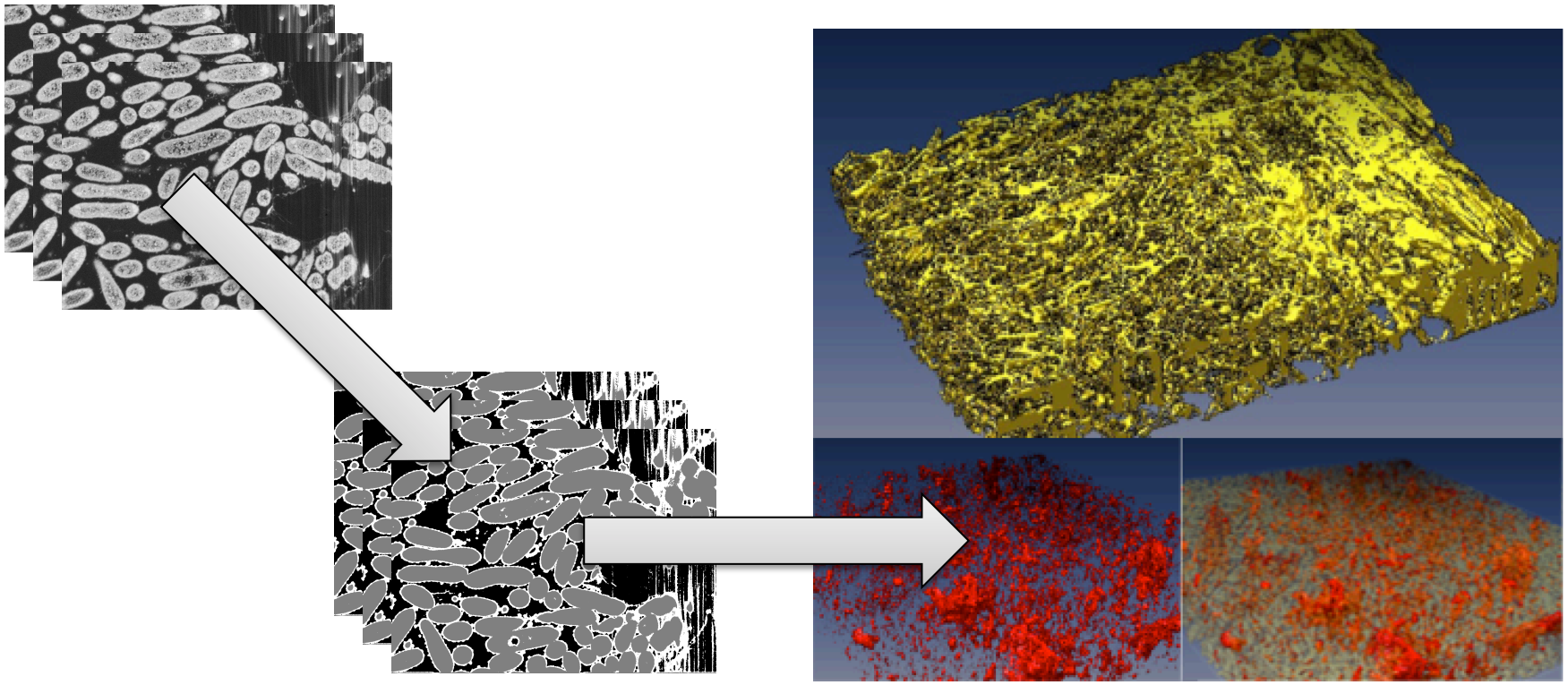
Results, ML & ImageJ-*based* segmentation



Microbial communities (*Desulfovibrio RCH1*) - Segmented metal precipitate

Results, ML-*based* segmentation

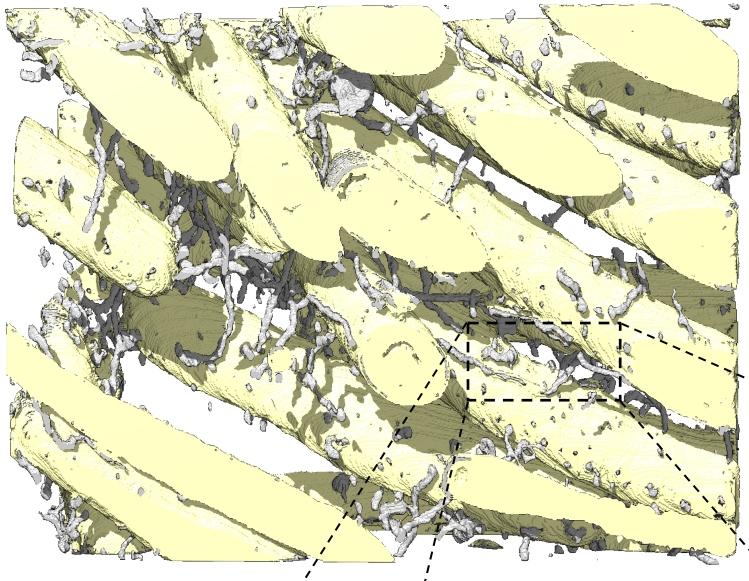
- **Segmentation of 3D images (5D capable)**

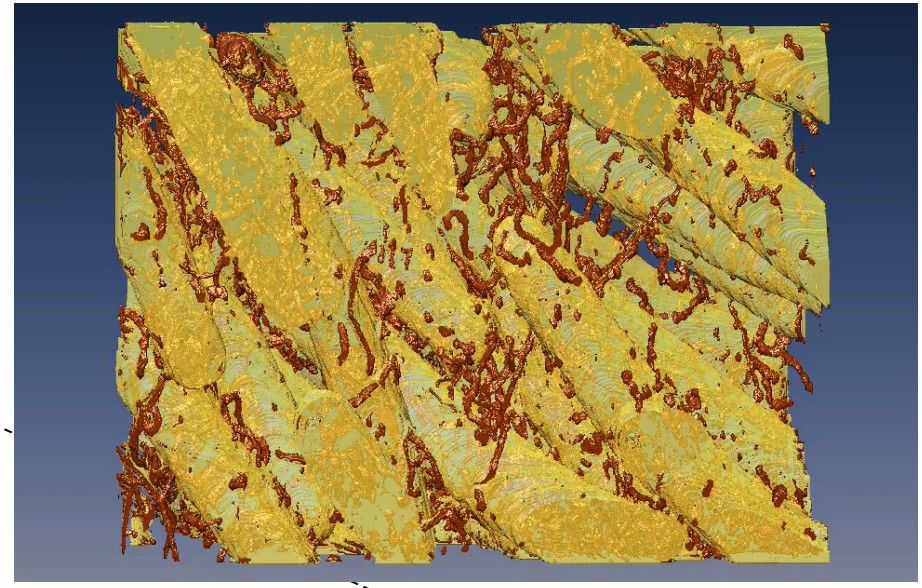


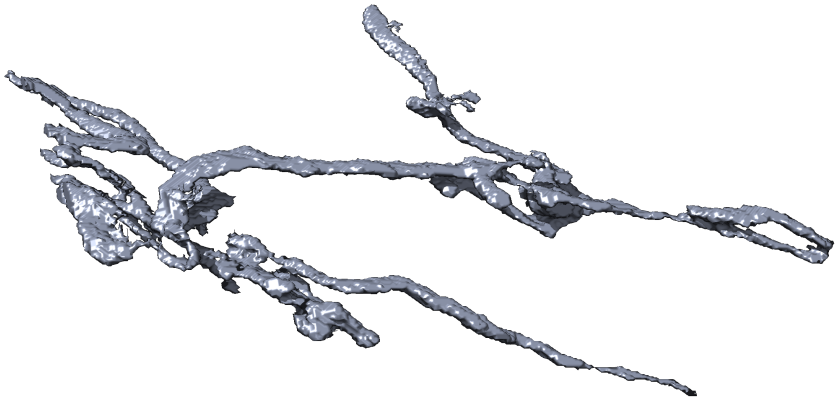
Un-processed MRC stack (800 images ~20MB each) → *Segmented maps (bacteria, background and intermediate features)* → *3D renderings of individual channels*

Segmented community (*Geobacter*)

Results, Big image vis





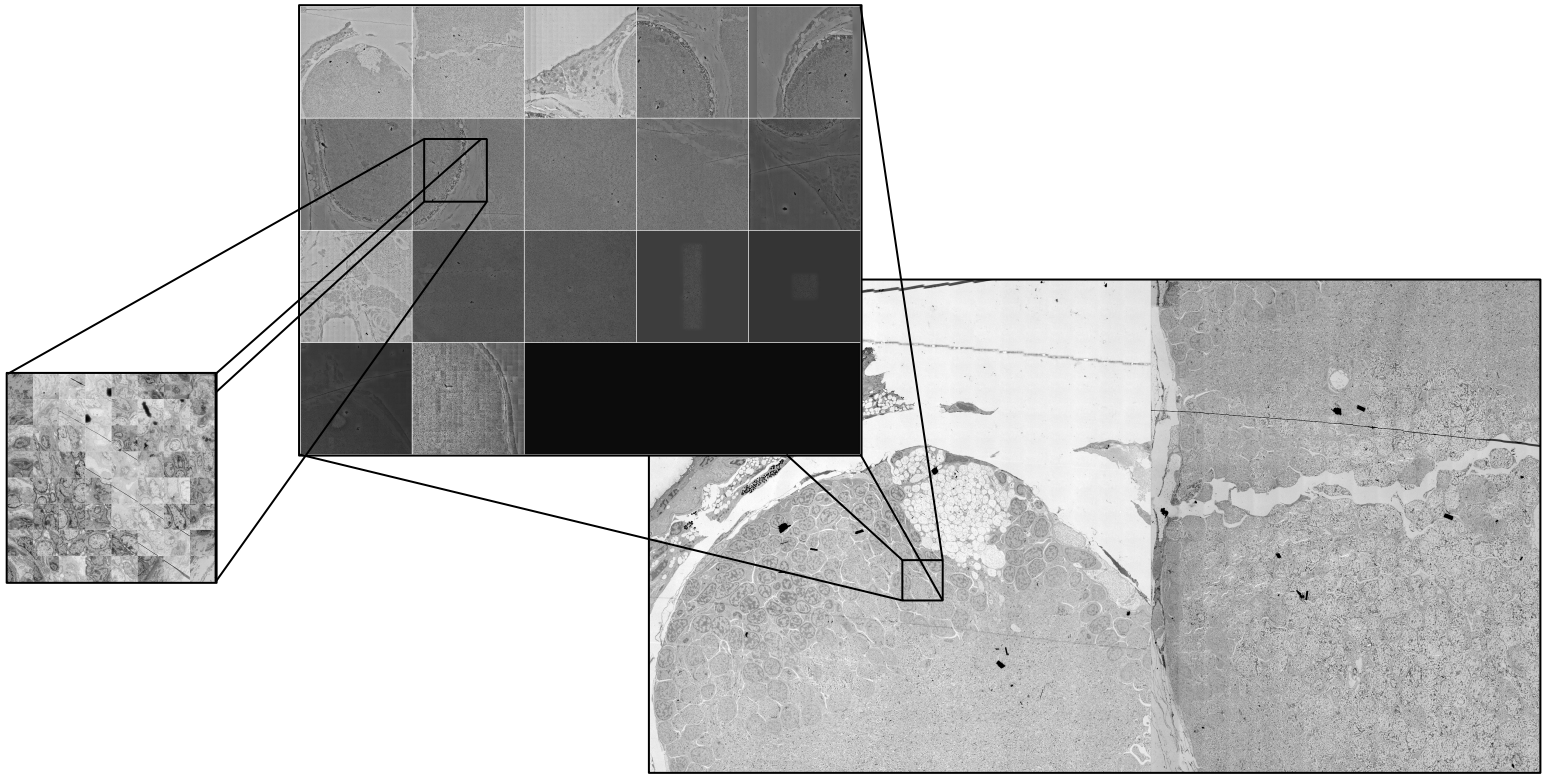


Visualizing both macromolecular details and community organization

Segmented community (*Myxococcus Xanthus*)

Results, High res. complex systems

- **Montaging 2D high-res images**



Un-processed, un-montaged MRC stack (64 images ~3GB) → *Processed, equalized and montaged single tiles (64 images x 375 per stack)* → *~ Gigapixel image dataset (800,000px by 600,000px) ~50GB*

Zebrafish head sample

Services

Services

- **Pre-processing**
  - Reconstruction
  - Filtering
  - …
- **Segmentation**
  - (Un)Supervised
  - Machine learning-based
  - …
- **Visualization**
  - 2D/3D renders
  - Patterns/Clusters
  - …
- **Analysis**
  - Descriptors
  - Geometries (NYI)
  - …

Software

- **Pre-processing**
  - IMOD-eTOMO/WBP
  - VLFEAT/PIL
  - Others
- **Segmentation**
  - Norm-based, Correspondence-based
  - NN/FL-based
  - Others
- **Visualization**
  - Plane/Volume vis (Chimera, ImageJ)
  - Mesh/Grid vis (NYI)
  - Others
- **Analysis**
  - Texture, feature, ROIs, location
  - Shape, angle, distance, volume
  - Others

Upcoming implementations

- Crowd sourced parameters for classification and segmentation
- Agile dev methodology engagement
- More Docs
- Public views for giga-pixel images and statistical information.

Thank you!