Workshop on Management of Large CryoEM Facilities

PROCESSING…

Edward T Eng | Anchi Cheng

February 6, 2017

SIMONS ELECTRON
MICROSCOPY CENTER

New York Structural
Biology Center

Processing at SEMC

Community

Infrastructure

Efficiency

Balance

Strategy

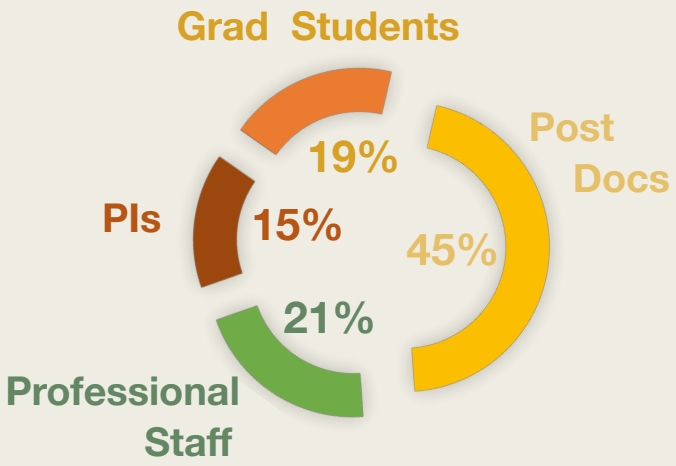
Community

User profile

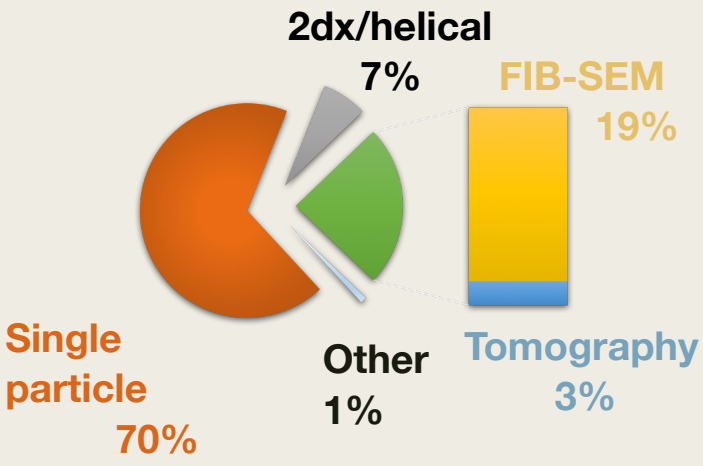
New York Structural Biology Center

Department of Health

Breakdown of users

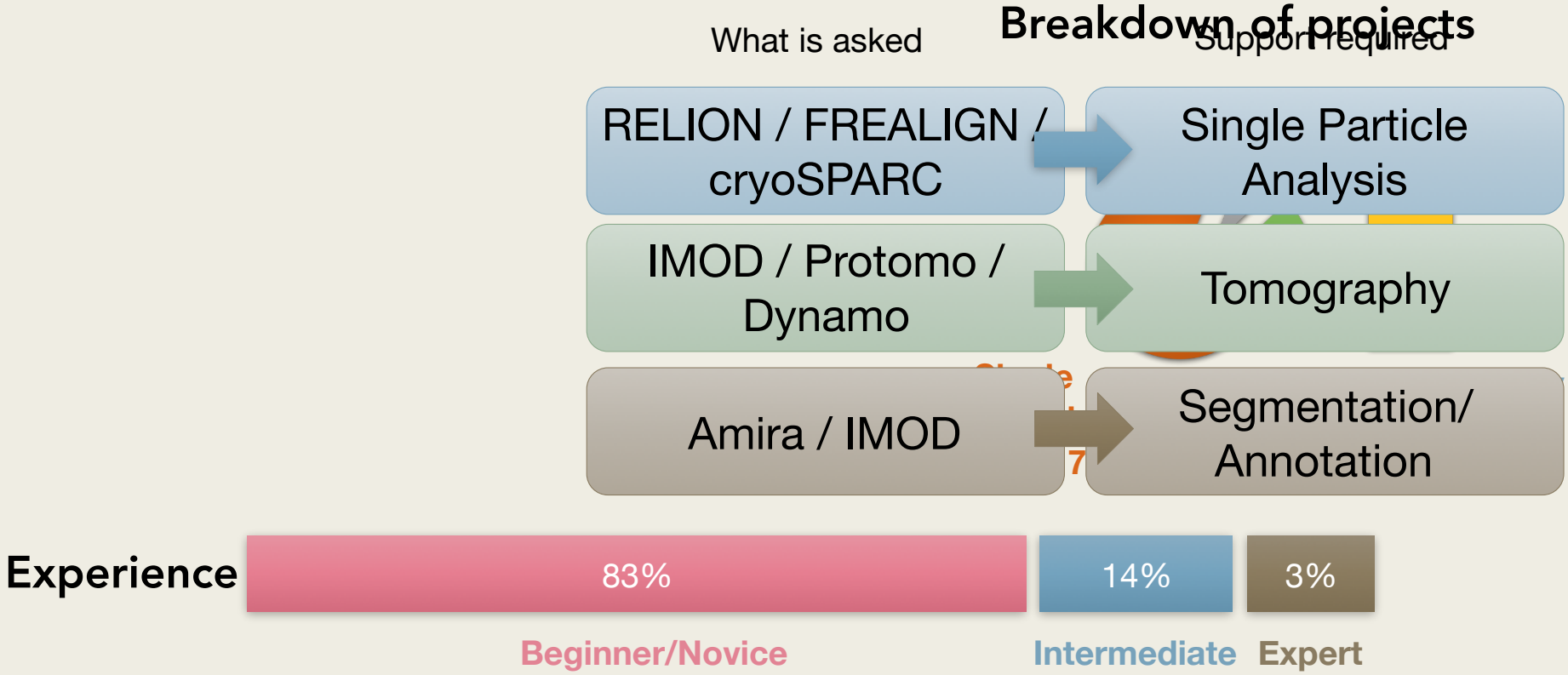


Breakdown of projects



Community

Top requests for processing



Infrastructure

2017 data generation at SEMC

EM instrumentation

FEI Titan Krios#1 / #2 / #3
Falcon3 x3
K2 x3

FEI Tecnai F20
DE20
TVIPS 4K CMOS

FEI Tecnai Biotwin
TVIPS 4K CMOS

JEOL 1230
Gatan US4000 CCD

FEI Helios 650
Quorum cryostage

Cameras

7 direct detectors on 4 TEMs

4 CMOS/CCDs on 3 TEMs + 1 SEM

Since 2015:
1,548,411 total images
(1.1 mil images since 2016)

For 2016:
170,933 Krios1 direct detector movies (exposure images)

Data acquisition and processing platforms

EM instrumentation

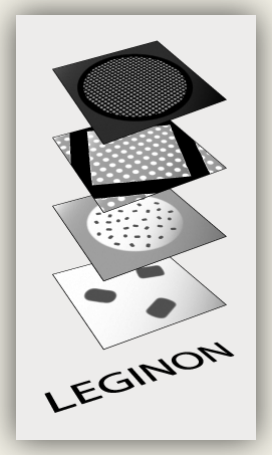
FEI Titan Krios#1 / #2 / #3
Falcon3 x3
K2 x3

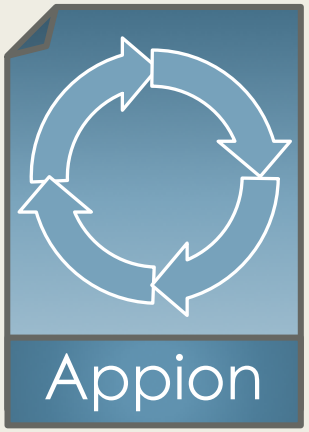
FEI Tecnai F20
DE20
TVIPS 4K CMOS

FEI Tecnai Biotwin
TVIPS 4K CMOS

JEOL 1230
Gatan US4000 CCD

FEI Helios 650
Quorum cryostage

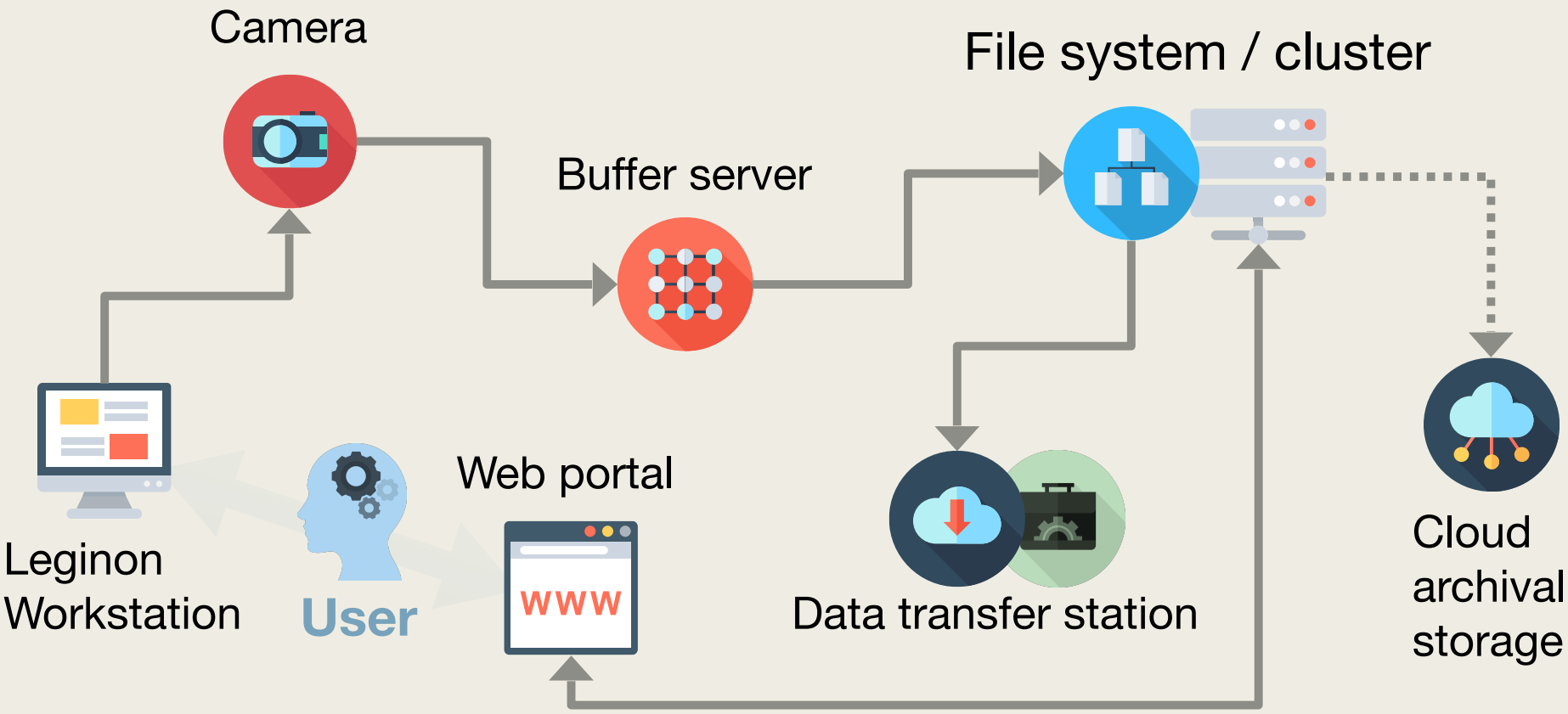




Infrastructure

On the fly data pipeline

c. 2017



SEMC computational support

- HPC Server and storage (DDN):
- 2 x 42U rack enclosures
- DDN GRIDScaler GS7K appliance with 1.1PB GPFS paralegal file system
- 420TB DDN WOS object storage for archival
- 1056 x CPU cores. 44 x SuperMicro nodes each with 24 x CPU cores and 256GB RAM
- 4 x GPU nodes each with one GPU and 128GB RAM. One GPU server with 8 x GPUs and 512GB RAM and 2 x GPU servers each with 4 x GPUs and 512GB RAM.
- 4 buffer servers each with 51TB local storage, 2 x GPUs, 128GB RAM and 10G Fiber Network cards.
- 5 x 36 QSFP port 56Gb/s FDR InfiniBand switches.
- Bright Cluster Manager
- Basic Onsite Support; 7x24 remote support

Additional computational resources

Web resources

- emgweb.nysbc.org
- emg.nysbc.org/ redmine
- semc.nysbc.org

Cluster resources

- SEMC Goby cluster
- SEMC Guppy cluster
- SEMC cluster

On site resources

EM session user

- Leginon / support workstations (9)
- GPU workstations (5)

Data processing user

- Amira workstations (2)
- Control room workstations (3)

- Data transfer computers (2)

Efficiency

Example: Single-particle workflow

During EM session

After EM session

Leginon session

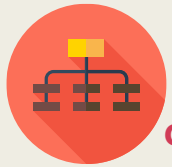
- Setting up workflow
- Data acquisition
- Workflow Optimization *if needed*

Appion session

- Frame alignment
- CTF estimation
- Particle picking
- Micrograph/ Particle curating

SEMC computing

- Micrograph/ Particle sorting
- Initial 2D classification
- Initial model generation
- 3D refinement



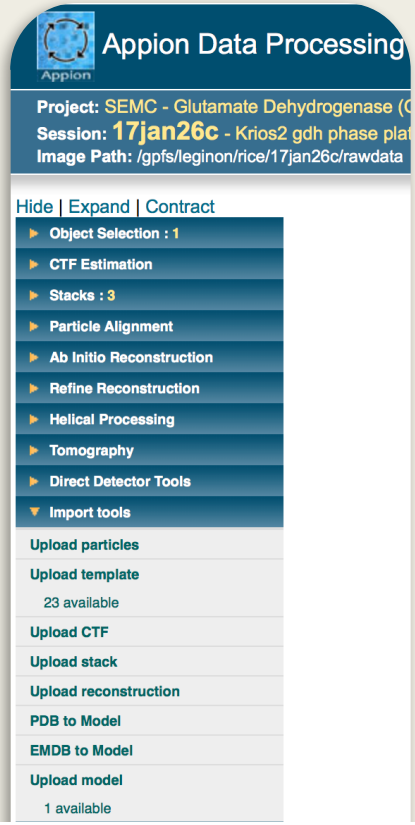
Home institution computing/cluster

- 2D classification
- 3D classification
- 3D refinement
- Model building

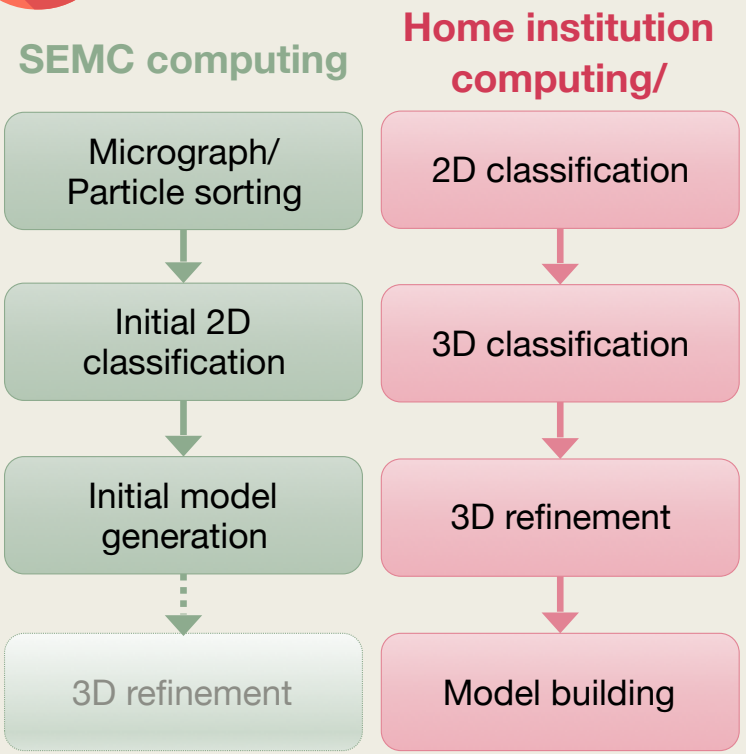
Efficiency

Example: Single-particle workflow

During EM session



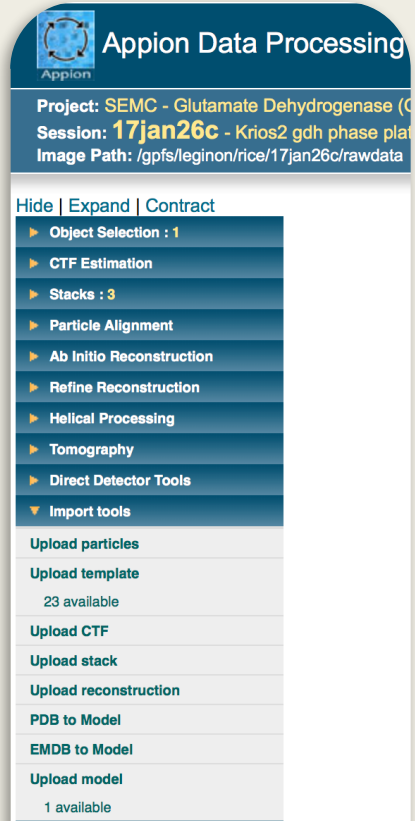
After EM session



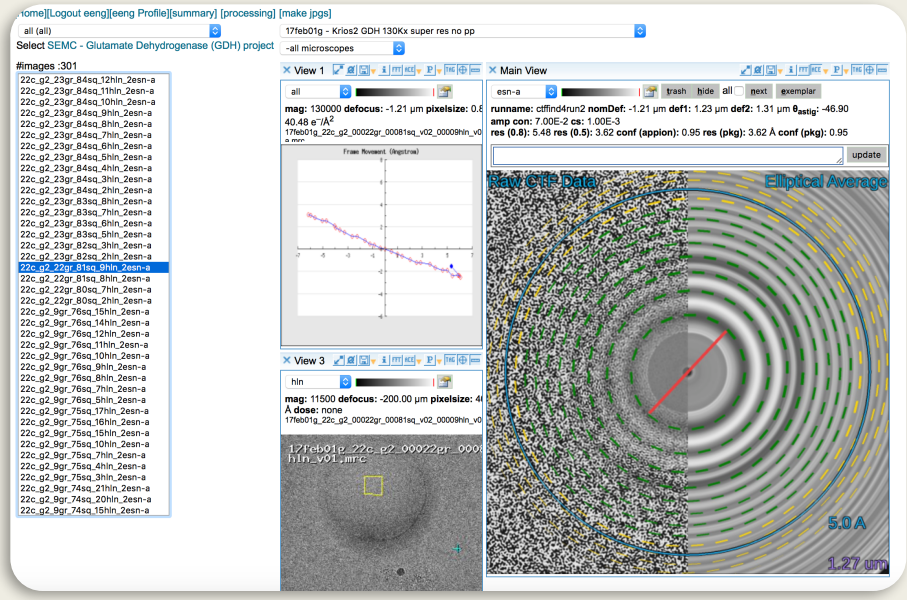
Balance

Web integrated processing

During EM session



Leginon / Appion web interface

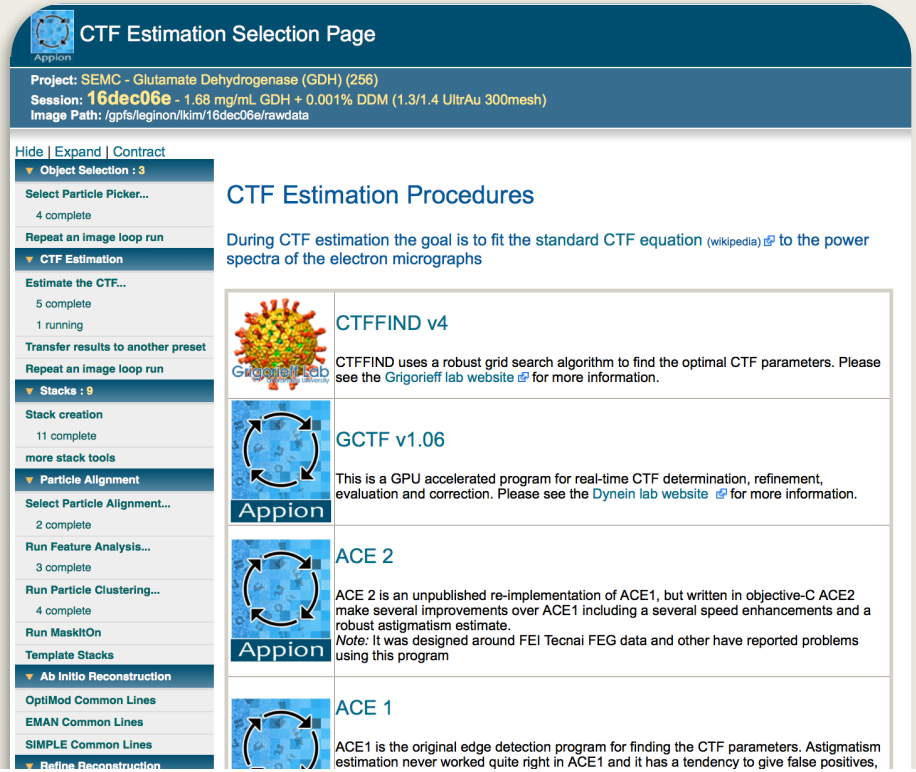


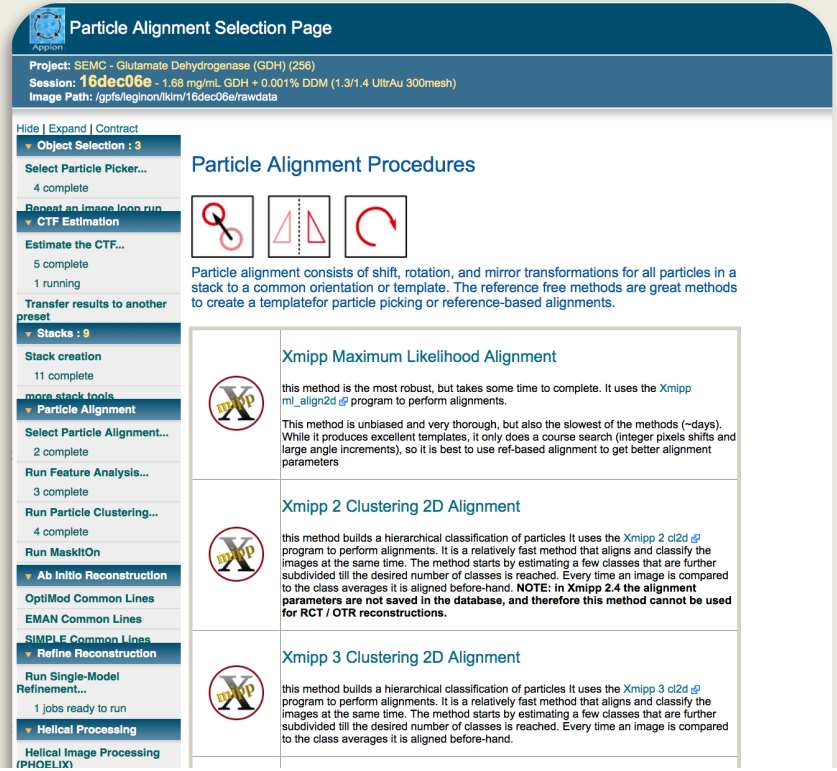
Balance

Packages offered

After EM session

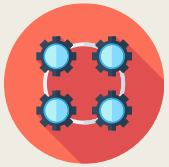
Appion web interface





Balance

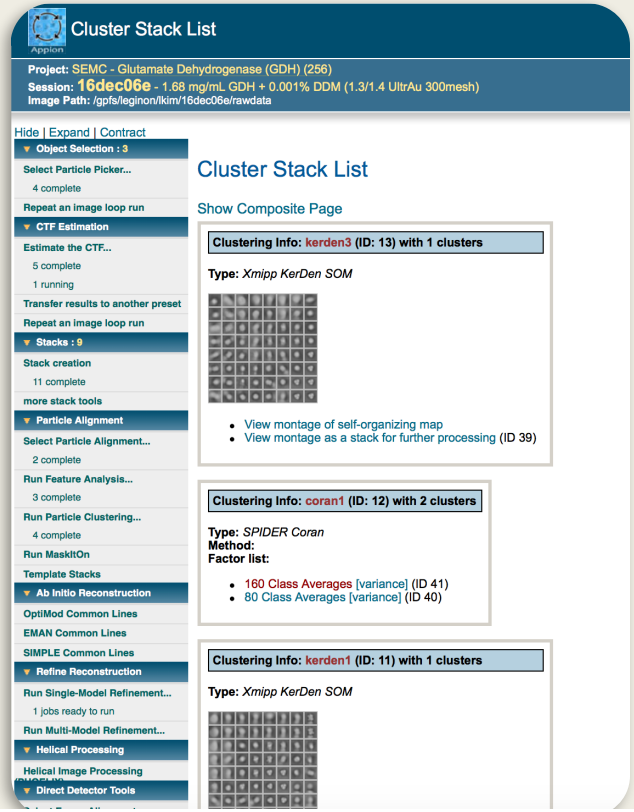
Who does the processing workflow?



After EM session

SEMC computing	Home institution computing/
Micrograph/ Particle sorting	2D classification
Initial 2D classification	3D classification
Initial model generation	3D refinement
3D refinement	Model building

Appion web interface



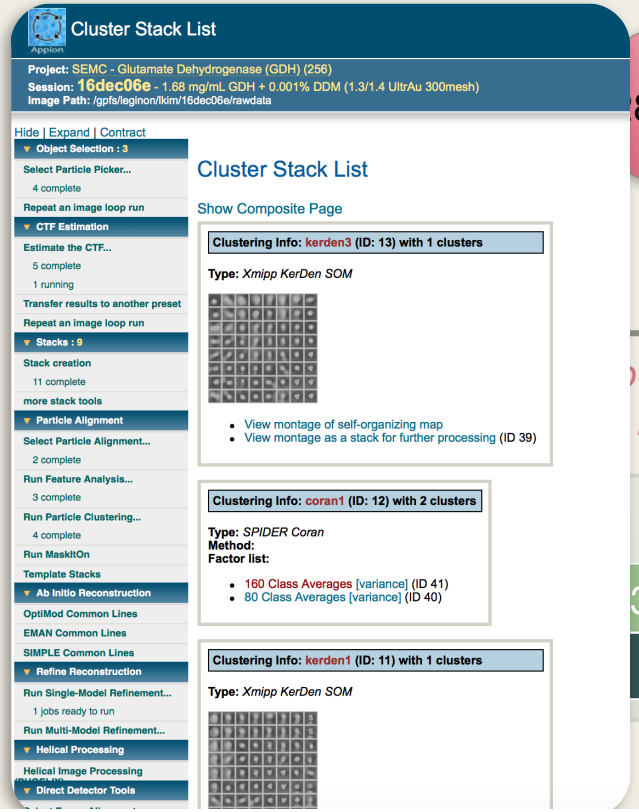
Balance

Who does the processing workflow?

After EM session

SEMC computing	Home institution computing/
Micrograph/ Particle sorting	2D classification
Initial 2D classification	3D classification
Initial model generation	3D refinement
3D refinement	Model building

Appion web interface for TEM use



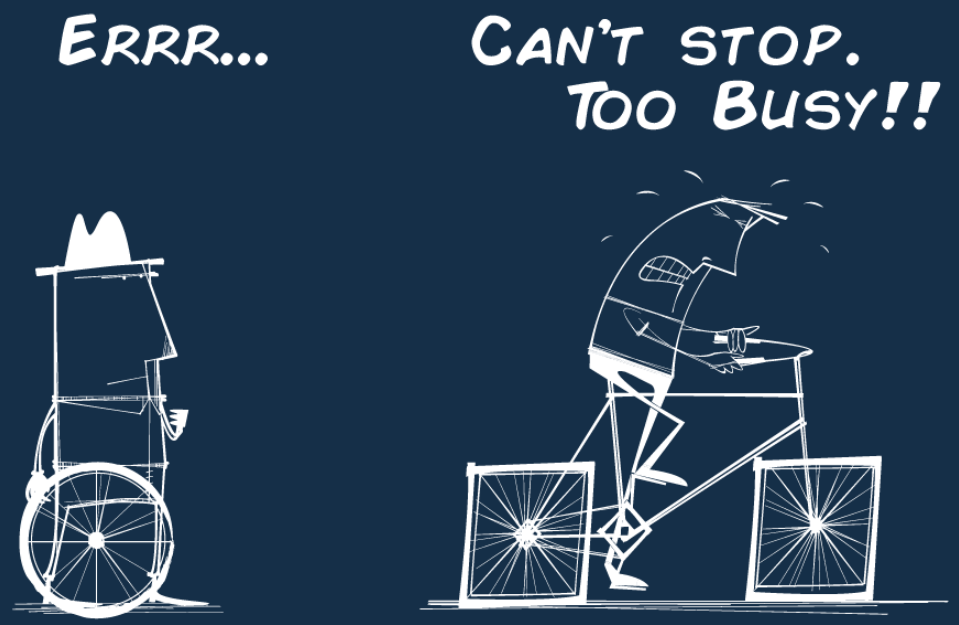
8,817

ocessing runs

3357

4855

Strategy

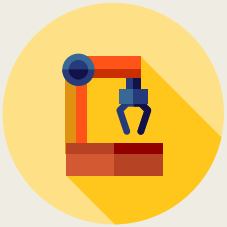


Too busy to improve?

Strategy

Feedback from the community

Development



Production

Time / Resources

Processing at SEMC

Community

What is our user community asking for?

Infrastructure

What is our infrastructure?

Efficiency

How does processing lead to more efficient data collection?

Balance

What is the balance between staff and user effort?

Strategy

How do we evolve and prepare for future advances?

PROCESSING
(cont)

before scaling up and support issues

…Anchi

Anchi Cheng
Res. Staff Scientist

Sargis Dallakyan
Res. Programmer

Carl Negro
Res. Programmer

Before 3 Krioses

EM instrumentation		Cameras
FEI Titan Krios#1 K2 x1		**2** direct detectors on 2 TEMs
FEI Tecnai F20 DE20 TVIPS 4K CMOS		
JEOL 1230 Gatan US4000 CCD		**3** CMOS/CCDs on 1 TEMs + 1 SEM
FEI Helios 650 Quorum cryostage		

2015-2016 computation support

- Guppy (8 nodes with 4-8 CPU cores)
- 1 Database and 1 web server
- 3 gpu workstations
- Individually mounted storage 300 TB

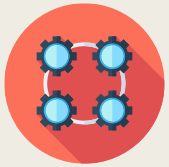
- GPFS storage 0.5 PB
- DE20 dedicated processing server (24 Processors)
- 1056 x CPU cores. 44 x SuperMicro nodes each with 24 x CPU cores and 256GB RAM
- 4 x GPU nodes each with one GPU and 128GB RAM.

Efficiency

Typical Problems

During EM session

- Frame alignment from the last user is not finished
- Web service image loading can not keep up with multiple user viewing.
- Database query lock up.
- Computation resource upgrade in progress



After EM session

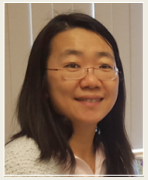
- Inefficient request of resource
  - *Ask more than needed*
  - *Ask less than required*
- Small one processor jobs compete with MPI processes.
- Users do not (know how to) clean up bad processing results.
- New development breaks the pipeline.

Simons Electron Microscopy Center

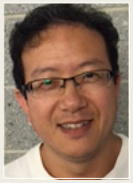
NEW YORK STRUCTURAL BIOLOGY CENTER
Simons Electron Microscopy Center


Anchi Cheng
Res. Staff Scientist


Sargis Dallakyan
Res. Programmer


Zhening Zhang
Res. Scientist


Carl Negro
Res. Programmer


Alex Wei
Technician


Phil Baldwin
Staff Scientist

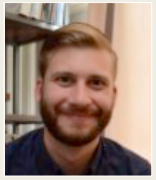

Clint Potter
Director


Bridget Carragher
Director


Yong Zi Tan
Grad. Student

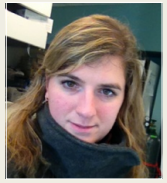

Venkat Dandey
Post Doc.


Alex Noble
Post Doc.


Micah Rapp
Grad Student


Priyamvada Acharya
Embedded Post Doc.


Giovanna Scapin
Embedded Scientist


Julia Brasch
Embedded Post Doc.

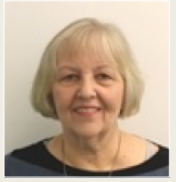

Bill Rice
EM Manager


Ed Eng
Staff Scientist


Ashleigh Raczkowski
Senior Technician


Kelsey Jordan
Technician


Laura Kim
Research Associate


Crystal Premo
Administrator

National Resource for Automated Molecular Microscopy
http://nramm.nysbc.org