HPC and Biomedicine

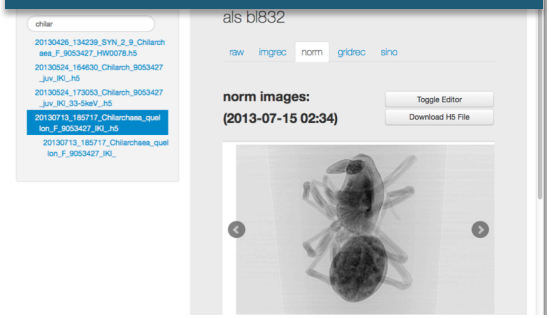
Kathy Yelick
Associate Laboratory Director
Computing Sciences
Lawrence Berkeley National Laboratory

Computing Capabilities of DOE

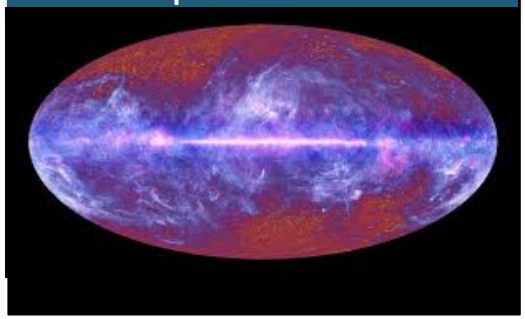
High Performance Systems

Physical Networks

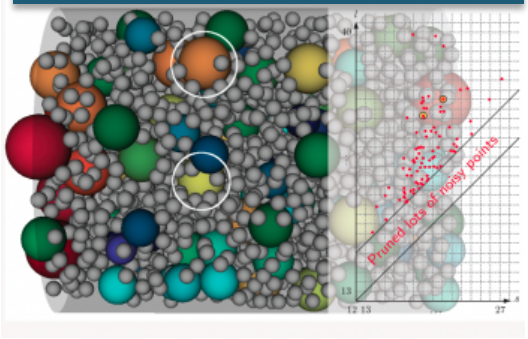
Software: Production development and support



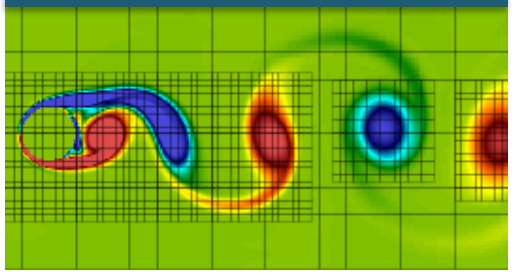
Culture of collaborative research on big science problems



Computer and Data Science, especially high performance



Applied Math



Collaboration models and example results

- **Institutes of expertise dedicated to Science of X**
  - CAMERA
- **Access to HPC systems and performance expertise**
  - HipMer
- **Long-term software and data infrastructure**
  - KBase
- **Co-developing instruments and analysis tools**
  - Brain and CryoEM
- **Grand challenges (shown throughout)**
  - Antibiotics
  - Cancer
  - Brain

Collaboration models and example results

- **Institutes of expertise dedicated to Science of X**
  - CAMERA
- **Access to HPC systems and performance expertise**
  - HipMer
- **Long-term software and data infrastructure**
  - KBase
- **Co-developing instruments and analysis tools**
  - Brain and CryoEM
- **Grand challenges (shown throughout)**
  - Antibiotics
  - Cancer
  - Brain

Center for Advanced Mathematics for Energy Research Applications (CAMERA)

Today:	Tomorrow:	Critical need:
Data analysis time-consuming	More data faster More resolution	algorithms/analysis for *understanding*

New math: Transform experimental data into understanding

- Extract information from murky data
- Interpret, and optimize experiments
- Deliver robust software tools
- Accelerate scientific discovery

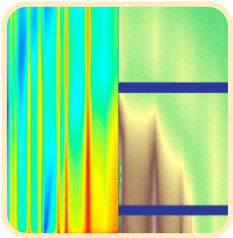
CAMERA Applied Math

James Sethian, PI

- Jointly funded by DOE ASCR (computing) and BES (light sources, materials and much more) after substantial internal funding

CAMERA: Mathematics for Experimental Science

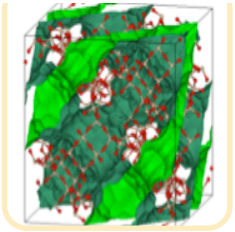
CAMERA Applied Math



GISAXS: X-ray scattering data analysis, 1000x faster

Science Drivers: BES User Facilities

Initially: ALS, LCLS, APS, NSLS, and MF



Material Informatics: E.g., Zeo++ high throughput porosity characterization

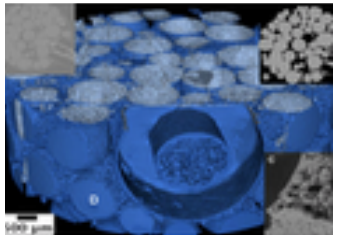
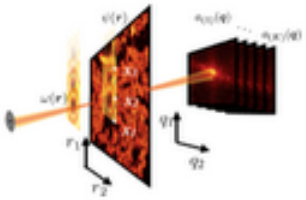
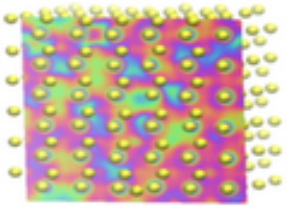


Image-Based Analysis: Automated Micro-CT sample analysis

X-ray Nano-Crystallography: solve image indexing problem



Ptychography: solve phase retrieval problem

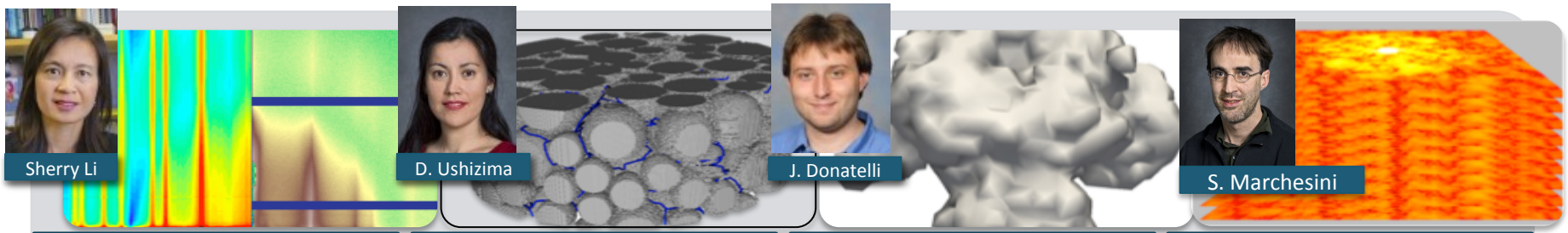


Electronic Structure: fast eigensolvers for materials

Foundation: state of the art mathematics

Spectral clustering, Maximum likelihood estimation, graph theory, machine learning, Mori-Zwanzig theory, Bilateral / anisotropic filters, PDE-based image segmentation, Computational harmonic analysis, Hamilton-Jacobi solvers, Bayesian analysis, Discrete Galerkin methods, Optimization methods

CAMERA leverages state-of-the-art mathematics to transform experimental data into understanding



X-ray scattering data analysis

400-1500x faster optimization

Now: Nonlinear optimization, genetic algorithms, pattern recognition w/ noise

Micro-CT Sample Analysis

Automated quantitative analysis

Now: 3D image segmentation; pattern recognition; classification; PDE- and graph-based analysis

X-ray Nano-Crystallographic Reconstruction

Indexing ambiguity resolved [PNAS13]

Now: Image orientation, find crystal shape/size; address orientation ambiguities; data variance reduction

Ptychography

Phase retrieval

Now: Provable convergence of algorithm; noisy data due to setup; select lens for specimen

Computing, experiments, networking and expertise in a "Superfacility" for Science

Collaboration models and example results

- Institutes of expertise dedicated to Science of X
  - CAMERA
- **Access to HPC systems and performance expertise**
  - HipMer
- Long-term software and data infrastructure
  - KBase
- Co-developing instruments and analysis tools
  - Brain and CryoEM
- **Grand challenges (shown throughout)**
  - Antibiotics
  - Cancer
  - Brain

De novo Genome Assembly

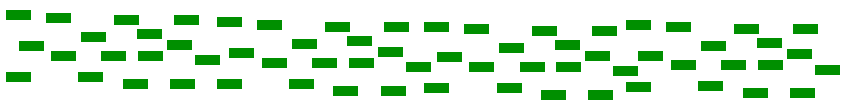
reads

Input: Reads that may contain errors

Chop reads into k-mers, process k-mers to **exclude errors**

1

***k*-mers**



Construct & traverse de Bruijn graph of k-mers, generate contigs

2

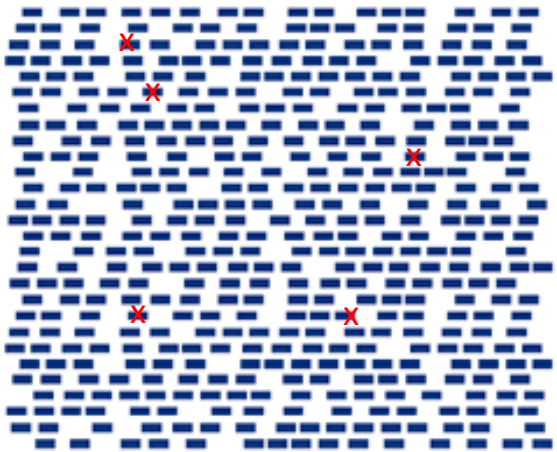
contigs

3

Use read information to link contigs and generate scaffolds.

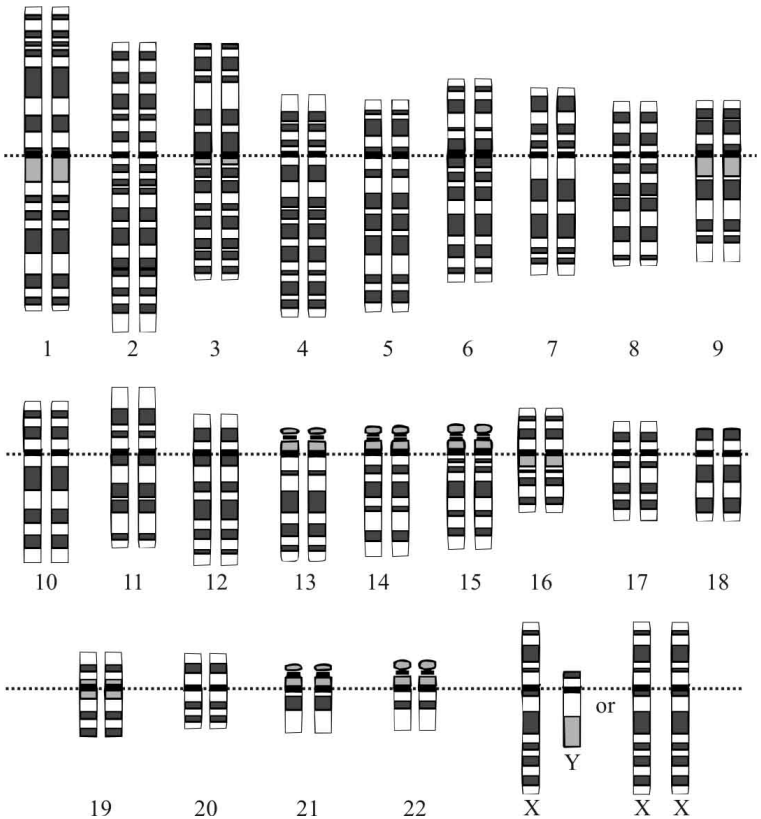
scaffolds

De novo Genome Assembly



>>10^9 sequencing reads

36 bp - 1 kb



3 Gb

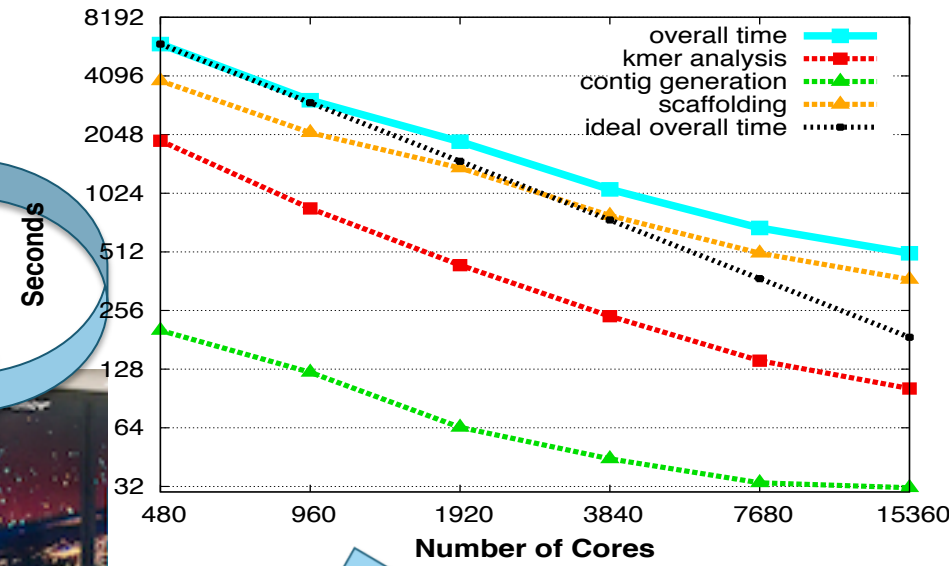
High Performance de Novo Genome Assembly

High Performance Meraculous assembler → HipMer

Computer Science HPC Expertise

- Remote Atomics
- Dynamic Aggregation
- Software Caching (sometimes)
- Clever algorithms and data structures (bloom filters, locality-aware hashing)
- Efficient languages (C vs Perl)
- Fast I/O

HPC systems with high speed interconnect networks



Grad student + software engineers needed to get to production

How Fast Is It?

HipMer = High Performance Meraculous assembler

- **Human genome (3Gbp):**
  - SGA assembler: 140 hours
  - Meraculous: 48 hours
  - HipMer: 8 minutes **(360x speedup)**

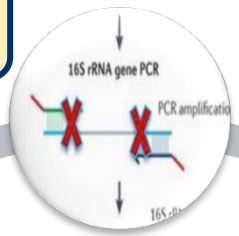
- **Wheat genome (17 Gbp):**
  - Meraculous (did not run, 170 hours projected):
  - HipMer: 39 minutes; 15K cores **(first all-in-one assembly)**

- **Wetland metagenome (1.25 Tbp):**
  - Meraculous (projected): 15 TB
  - HipMER: 11 minutes; 20K cores **(contig generation)**

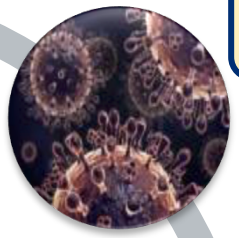
Metagenomics data mining efforts at JGI

Assembled metagenomes

Blind Spots
Nature Microbiol
2016

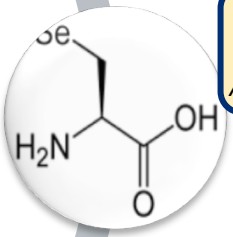


Earth Virome
Submitted



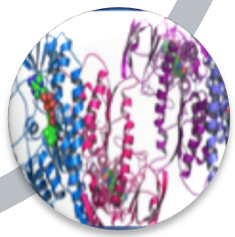
Selenocysteine Recoding
A. Chemie in press

With
D. Soll
Yale Univ.



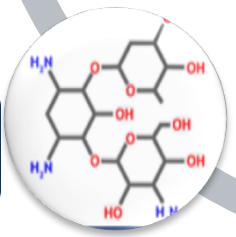
Novel Protein folds
in preparation

With
D. Baker
U. Washington



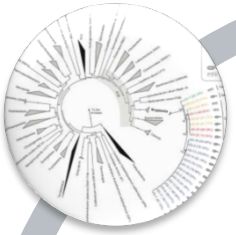
Metagenomic Protein clusters
in progress

Biosynthetic clusters
in progress



Recoding
Science, 2014

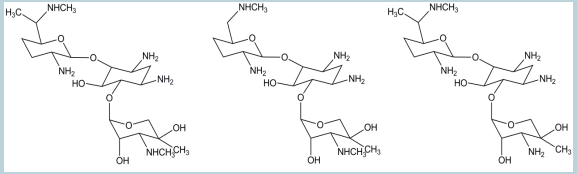
Kryptonia
Nature Com.
2016



JGI

Biofoundry: Rapid Production of Antimicrobials

New antibiotic-resistant pathogen



Screen drug variants for efficacy

Stockpiled vials of cells to produce drug variants

Distributed fermentation drug production facilities

Rapid surge production of effective drug variant

Grand Challenge:

- Discover new and improved antimicrobials for human, animal, and plant pathogens
- Rapidly identify an effective antibiotic and surge its production at distributed sites

Collaboration models and example results

- Institutes of expertise dedicated to Science of X
  - CAMERA
- Access to HPC systems and performance expertise
  - HipMer
- **Long-term software and data infrastructure**
  - KBase
- Co-developing instruments and analysis tools
  - Brain and CryoEM
- Grand challenges (shown throughout)
  - Antibiotics
  - Cancer
  - Brain

KBase: DOE Systems Biology Knowledgebase

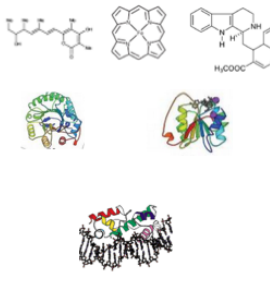
Open software and data platform for addressing the grand challenge of systems biology: **Predicting and designing biological function**

Unified system that integrates data and analytical tools for comparative functional genomics of **microbes, plants, and their communities**

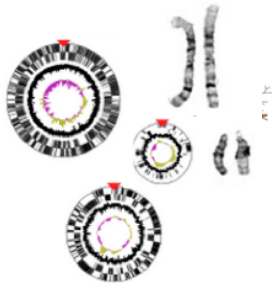
Collaborative environment for **sharing methods and results** and placing those results in the context of knowledge in the field

KBase: DOE Systems Biology Knowledgebase

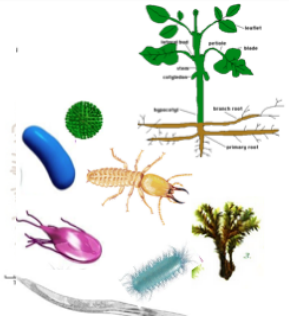
KBase Scope of Operations



Biomolecular Mechanisms, Models and Networks



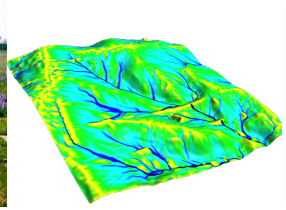
Genomics and Functional Genomics



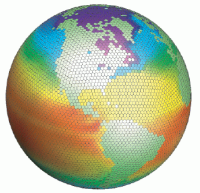
Organismal Biology, Dynamics and Interactions

Biodesign & Pore-scale dynamics and Biotic-Abiotic interactions

BioGeoMolecular dynamics, Trait-based models, Biogeochemical Cycling



Terrestrial Ecology and Subsurface Biogeochemistry



Watershed Hydrobio, Dynamic Vegetation observation and models

Land atmosphere, Earth System observation and models

- **Data: 28,300 genomes, 36,700 metabolic models, 27,000 compounds, 33,000 Reactions, 520 media types**
- **Tools: assembly, annotation, analysis, comparison to models**

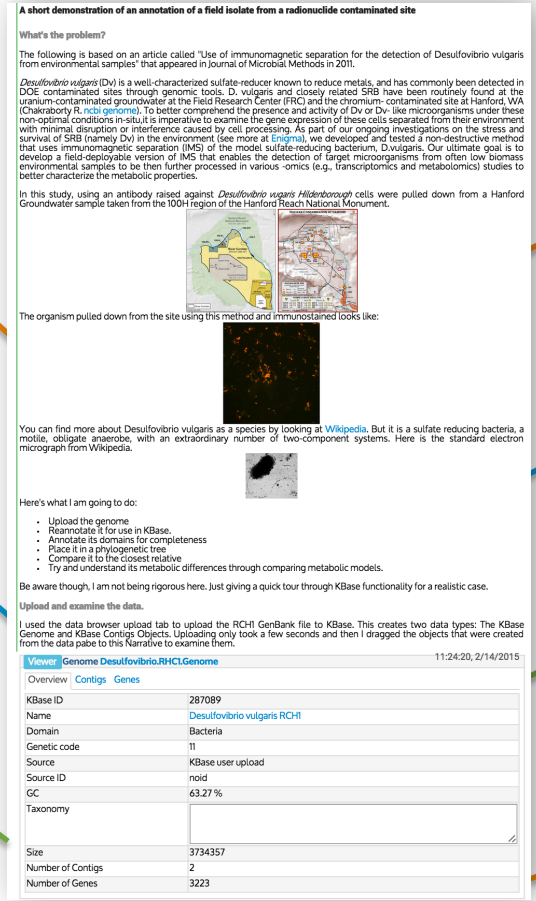
“Narrative interface” for collaborative science

An interactive, dynamic, and persistent document created by users that promotes open, reproducible, and collaborative science

Data

Analysis steps

Version control and provenance



A short demonstration of an annotation of a field isolate from a radionuclide contaminated site

What's the problem?

The following is based on an article called "Use of immunomagnetic separation for the detection of Desulfovibrio vulgaris from environmental samples" that appeared in Journal of Microbial Methods in 2011.

Desulfovibrio vulgaris (Dv) is a well-characterized sulfate-reducer known to reduce metals, and has commonly been detected in DOE contaminated sites through genomic tools. D. vulgaris and closely related SRB have been routinely found at the uranium-contaminated groundwater at the Field Research Center (FRC) and the chromium- contaminated site at Hanford, WA (Chakraborty R, ncbi genome). To better comprehend the presence and activity of Dv or Dv- like microorganisms under these non-optimal conditions in-situ,it is imperative to examine the gene expression of these cells separated from their environment with minimal disruption or interference caused by cell processing. As part of our ongoing investigations on the stress and survival of SRB (namely Dv) in the environment (see more at Enigma), we developed and tested a non-destructive method that uses immunomagnetic separation (IMS) of the model sulfate-reducing bacterium, D.vulgaris. Our ultimate goal is to develop a field-deployable version of IMS that enables the detection of target microorganisms from often low biomass environmental samples to be then further processed in various -omics (e.g., transcriptomics and metabolomics) studies to better characterize the metabolic properties.

In this study, using an antibody raised against *Desulfovibrio vulgaris Hildenborough* cells were pulled down from a Hanford Groundwater sample taken from the 100H region of the Hanford Reach National Monument.

The organism pulled down from the site using this method and immunostained looks like:

You can find more about Desulfovibrio vulgaris as a species by looking at Wikipedia. But it is a sulfate reducing bacteria, a motile, obligate anaerobe, with an extraordinary number of two-component systems. Here is the standard electron micrograph from Wikipedia.

Here's what I am going to do:

- Upload the genome
- Reannotate it for use in KBase.
- Annotate its domains for completeness
- Place it in a phylogenetic tree
- Compare it to the closest relative
- Try and understand its metabolic differences through comparing metabolic models.

Be aware though, I am not being rigorous here. Just giving a quick tour through KBase functionality for a realistic case.

Upload and examine the data.

I used the data browser upload tab to upload the RCH1 GenBank file to KBase. This creates two data types: The KBase Genome and KBase Contigs Objects. Uploading only took a few seconds and then I dragged the objects that were created from the data pabe to this Narrative to examine them.

11:24:20, 2/14/2015

Viewer Genome Desulfovibrio.RHC1.Genome

Overview Contigs Genes

KBase ID	287089
Name	Desulfovibrio vulgaris RCH1
Domain	Bacteria
Genetic code	11
Source	KBase user upload
Source ID	noid
GC	63.27 %
Taxonomy	
Size	3734357
Number of Contigs	2
Number of Genes	3223

Commentary

Visualizations

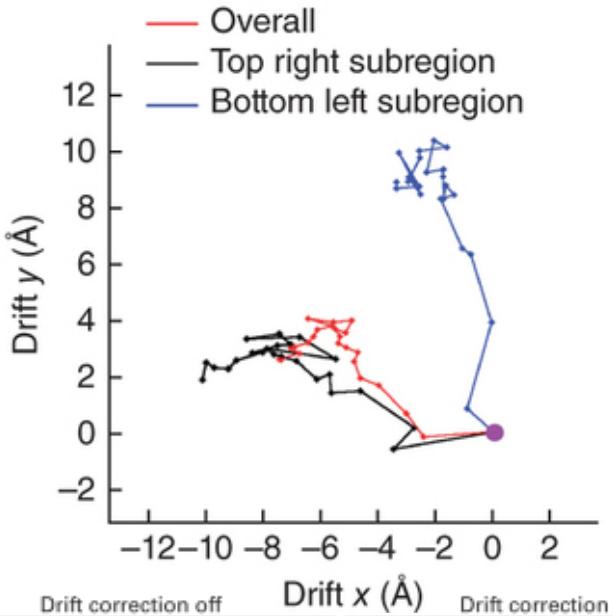
Custom scripts

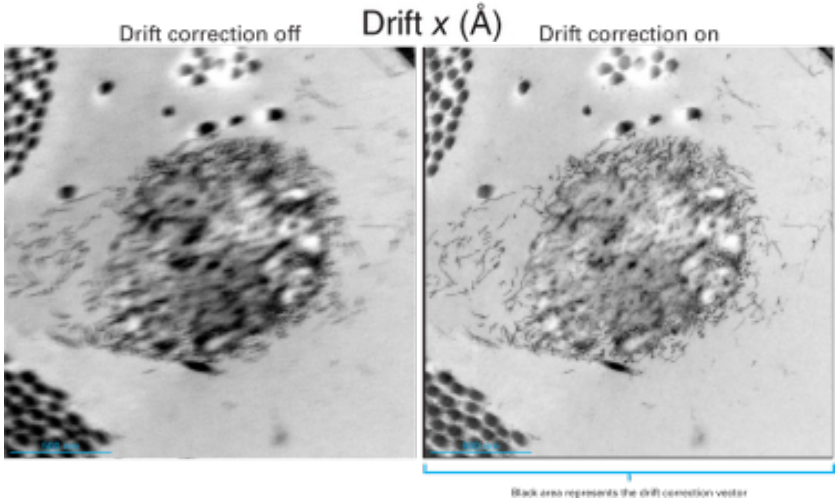
Sharing

Collaboration models and example results

- Institutes of expertise dedicated to Science of X
  - CAMERA
- Access to HPC systems and performance expertise
  - HipMer
- Long-term software and data infrastructure
  - KBase
- **Co-developing instruments and analysis tools**
  - Brain and CryoEM
- **Grand challenges (shown throughout)**
  - Antibiotics
  - Cancer
  - Brain

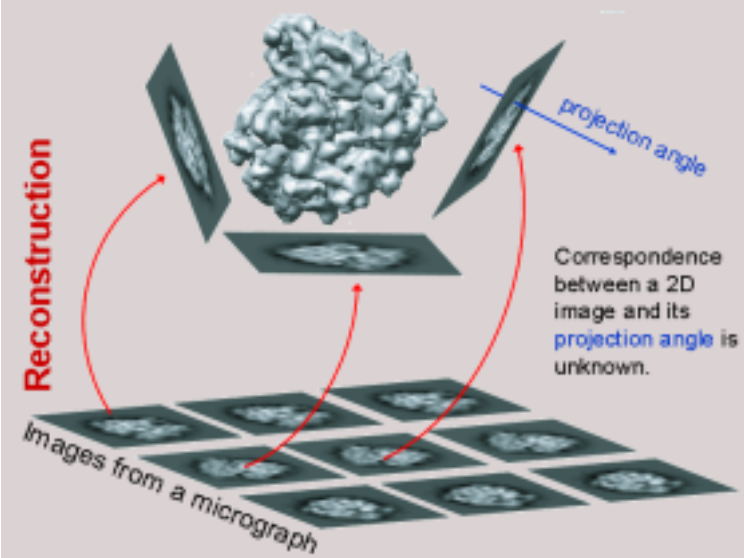
Impact of Direct Detectors (DOE developed)





- **DEDs have higher sensitivity and resolution than film or CCDs**
  - New technologies being developed
- **Biggest advance is the rate of data acquisition (movies)**
- **The movies can be analyzed to correct for the particle movement caused by the electron beam**
- **Computing:**
  - Fast data rate, and large data storage
  - Real time computing for corrections

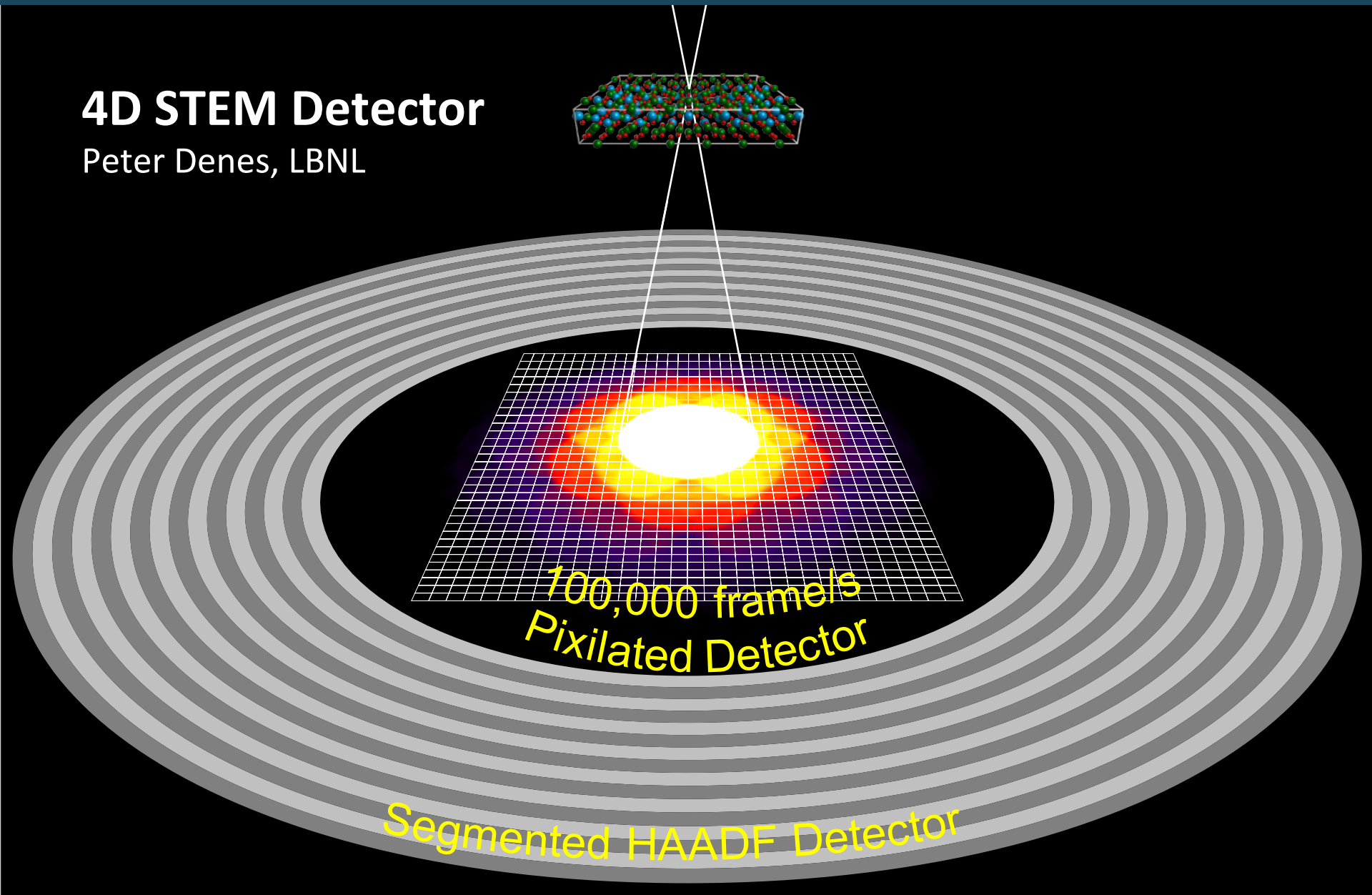
Cryo-EM Computational Issues



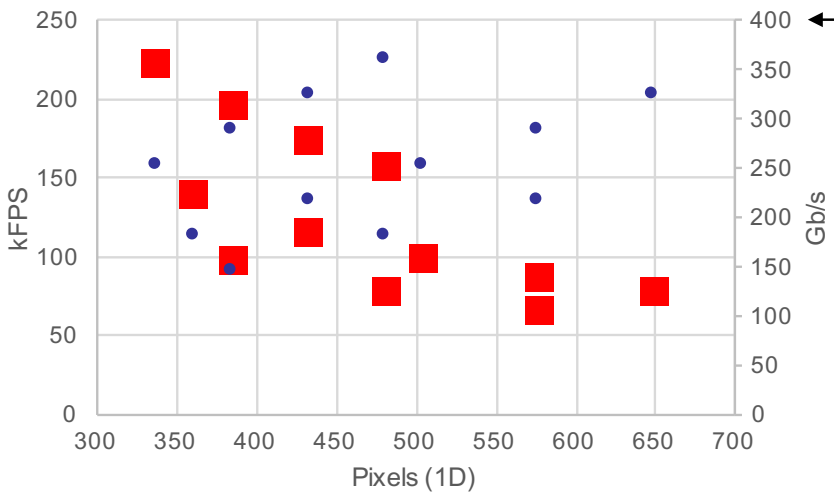
- **Many 2D projections of the 3D object need to be aligned to create a 3D reconstruction**
- **Many images must be held in memory (32-64GB per core)**
- **Current algorithms do not scale well**
- **Current codes do not scale well**

Current best practice is the use of Bayesian methods (RELION) and a single high resolution reconstruction will use 100-200 thousand particles and ~two weeks of 200-300 cores running in parallel

Technology is getting better; computation getting harder



Superfacility for 100,000 FPS Detector (for BES / DOE)



- Brocade: 400 Gb/s

Brocade
130 Holger Way, San Jose, CA 95134
T. 408.333.8000 F. 408.333.8101
www.brocade.com

BROCADE

April 7, 2015

Mr. Brent Draney
LBL-NERSC
415-20th Street
Oakland, CA 94612

Dear Brent,

Brocade has a long history of innovation and collaboration in the high tech research community. Continuing this tradition, Brocade would be honored to partner with NERSC on the "Future Electron Scattering Project" by loaning switching hardware. Brocade agrees to loan a switching layer for the project which provides at least ten ports of 40 gig and 4 ports of 100G by Q42016.

Brocade understands that at the end of the project all equipment will be returned to Brocade.

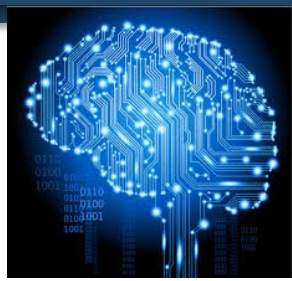
Sincerely,

Michael Bushong
Vice President
Data Center Switching and Routing

- 100 kFPS → 10s of TB / hour
- Real time analysis:
  - Sparsification
  - Clustering
  - Dedicated network to NERSC

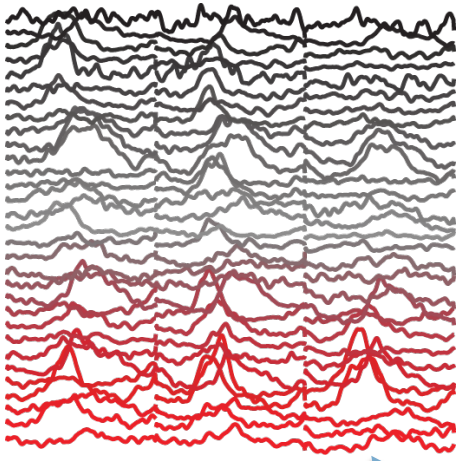
Possible DOE computing role in BRAIN

DOE can play a unique role in BRAIN computing through advances in applied mathematics and computer science together with HPC facilities.



Function

dynamic data

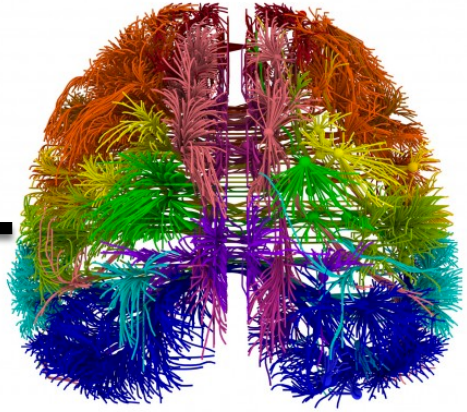


Theory & Models

abstractions

Structure

static data



Generation and analysis of raw data

Linking structure to function is a 'grand challenge' in general biology and materials

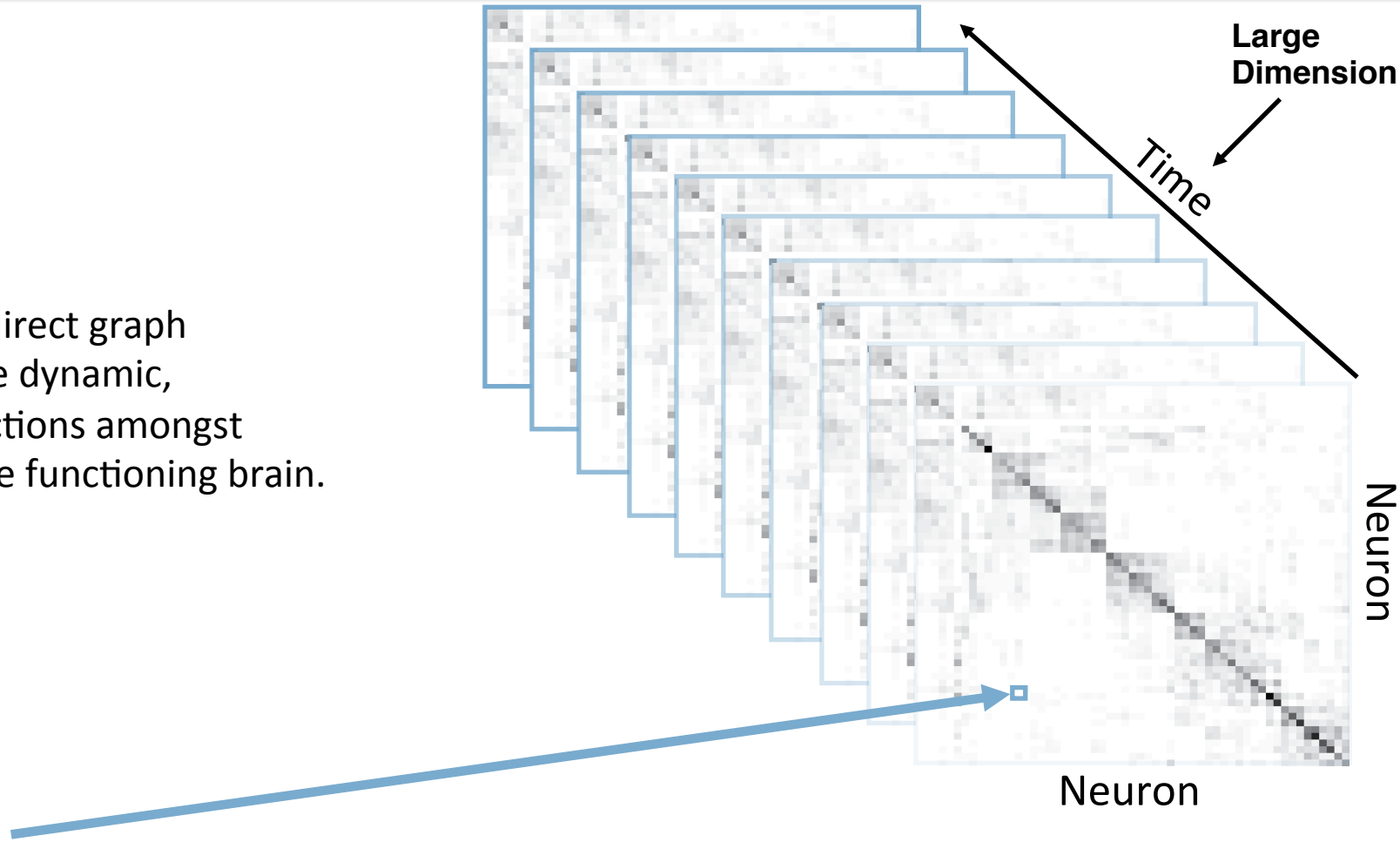
	-omics	Structure imaging	optical probes	electrical probes	optogenetics+	whole brain technologies	computing
1) Discovering diversity: provide access to **different brain cell types** to determine roles	F						
2) Maps at multiple scales: generate circuit diagrams that **vary in resolution**		TFF					ADF
3) Brain in action: Produce a **dynamic picture** of the functioning brain			TF	TF			ADF
4) Demonstrating causality: link brain activities to **intervention tools**					TF		
5) Identifying fundamental principles: develop **theoretical and data analysis tools**							MIA DF
6) Advancing human neuroscience: develop **technologies to understand human brain**				T		T	ADF
From BRAIN initiatives to the brain: Integrate technologies from 1-6 to **understand the brain and treat** disorders							

A Analysis
D Data
F Facilities
M Modeling
I Integration

F Facilities
T Tool development
F Facilities for tool development

The Functional Connectome

A weighted, direct graph describing the dynamic, casual interactions amongst neurons in the functioning brain.

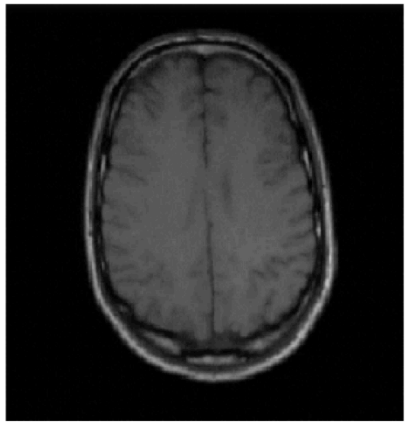


(e.g.) Each edge is **estimated** from data as partial correlation coefficient using **regression**.

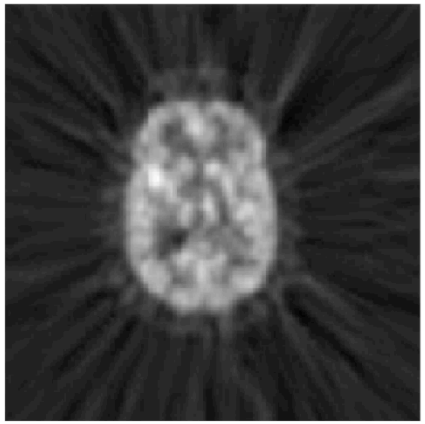
Source: K. Bouchard, LBNL

Multimodal Brain Analysis

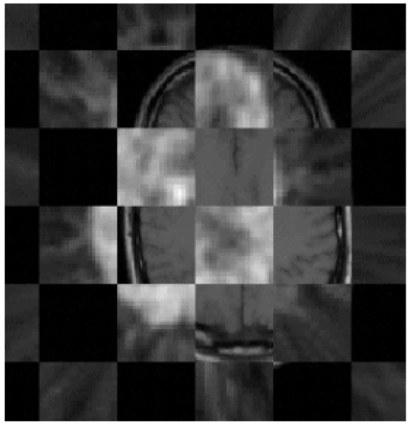
collaboration between UCB, LBNL, UCSF



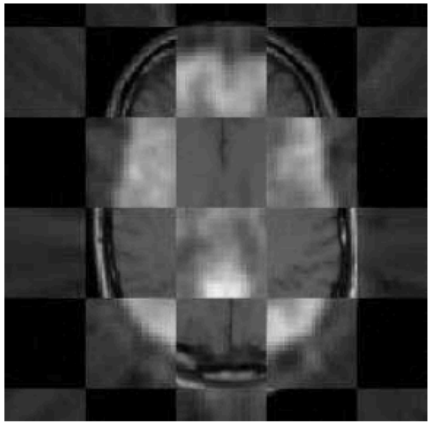
MRI



PET



Before registration



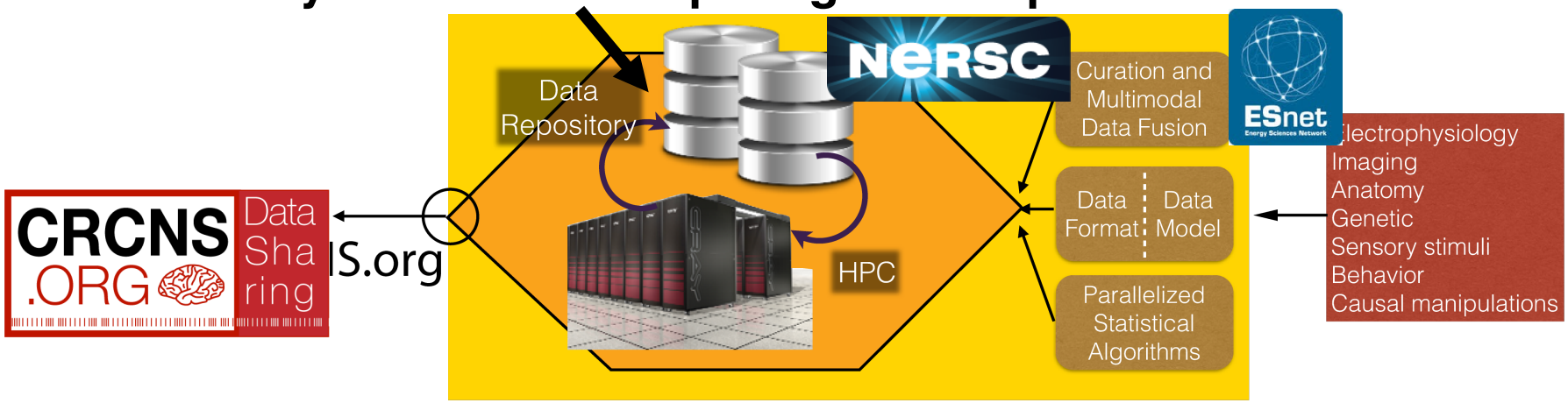
After registration

- **MRI with PET or cytology**
- **Optimization to find the spatial mapping to align images**
- **Linear algebra (SVD, LLS)**

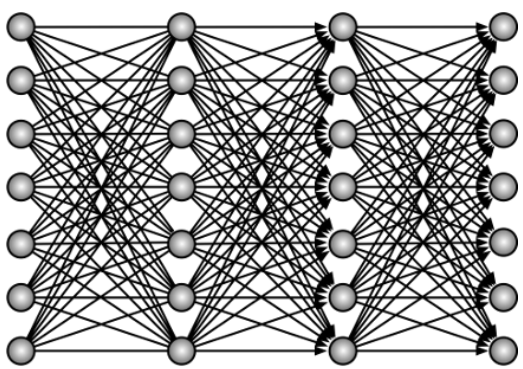
Source: S. Rit "Image registration", CREATIS Lab.

Advanced Computing for BRAIN

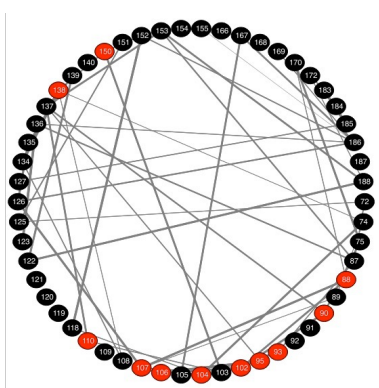
Familiar cycle in DOE computing: will require Exascale in 2025

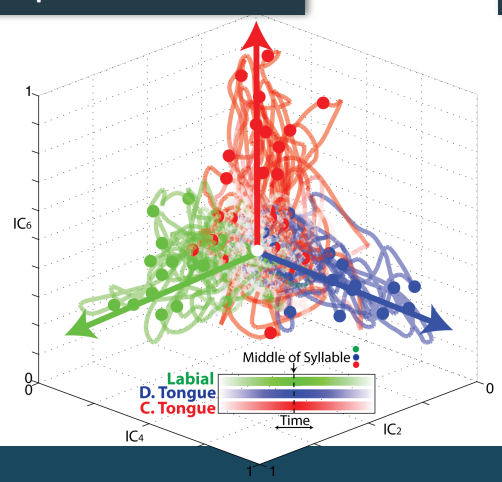


Deep neural networks for decoding brain activity

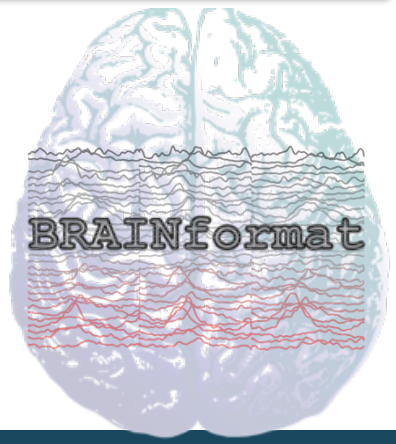


Sparse neural activity for human speech production





HDF5 format and data model for HPC



Source: K. Bouchard, LBNL

Machine learning for BRAIN are ubiquitous in DoE

DOE domains overlapping with methods for neuroscience

Methods for neuroscience overlapping with DOE domains

	Astronomy	Cosmology	Climate	Systems Biology	Neuroscience	BioImaging	Mass-spec	Personalized Medicine	Materials	Particle Physics
Classification	X		X		X*	X	X			X
Regression					X*			X	X	
Clustering		X	X		X		X			X
Dim. Reduction			X		X*		X			
Inference	X						X			X
Model Estimation	X				X*			X		
Image Processing	X					X				
Semantic Analysis			X	X					X	
Feature Learning			X		X		X	X	X	X
Anomaly Detection	X		X							X

*: discussed here

Source: Prabhat, NERSC

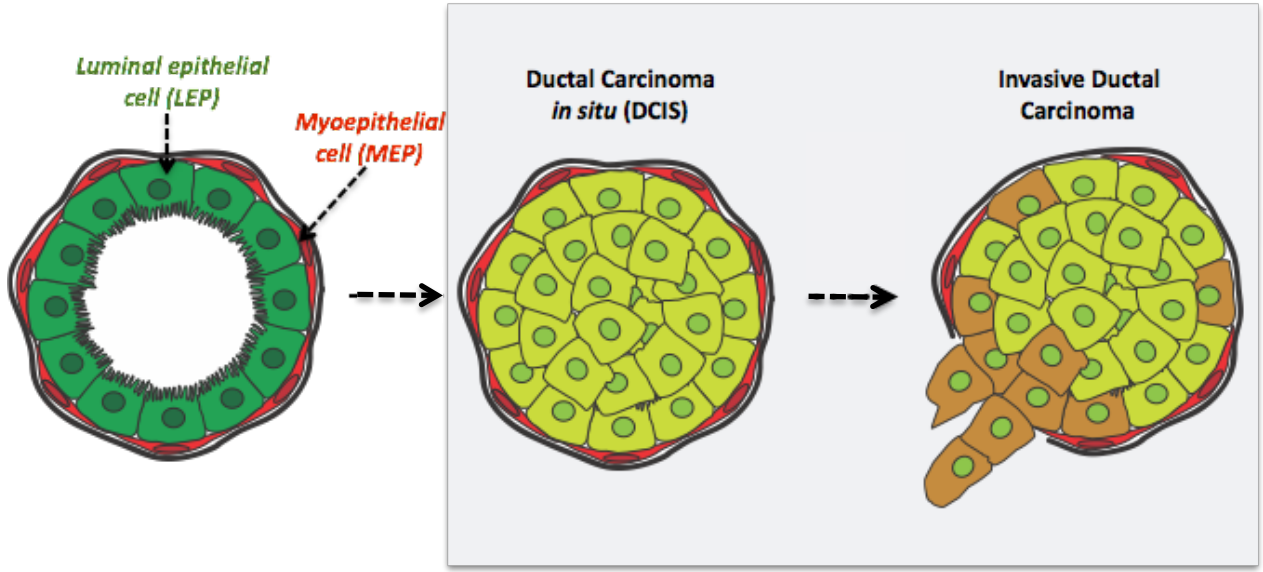
Collaboration models and example results

- Institutes of expertise dedicated to Science of X
  - CAMERA
- Access to HPC systems and performance expertise
  - HipMer
- Long-term software and data infrastructure
  - KBase
- Co-developing instruments and analysis tools
  - Brain and CryoEM
- **Grand challenges (shown throughout)**
  - Antibiotics
  - Cancer
  - Brain

Models to enable a detailed dissection of progression

Systems for studying the relationship between form (tissue architecture), function, and genetic information

Phenotypic stages



Molecular Stages

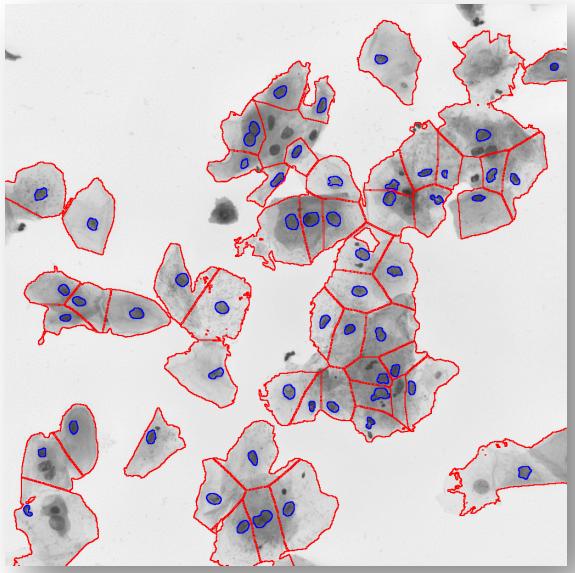
Confounding heterogeneity and passenger errors

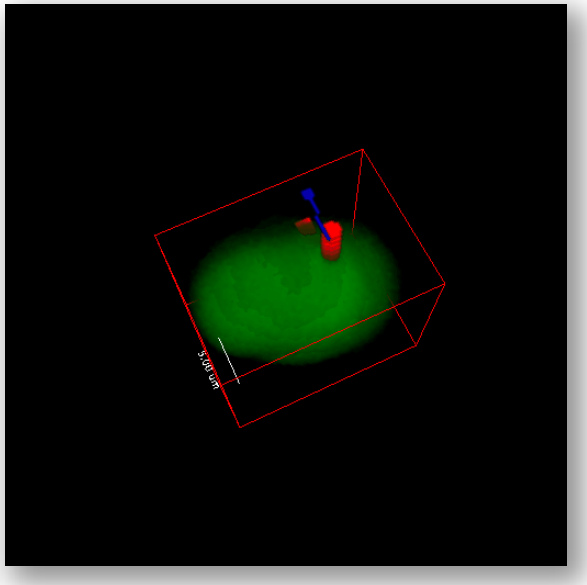
Pre-stasis → **Post-stasis** → **Immortal** → **Malignant**

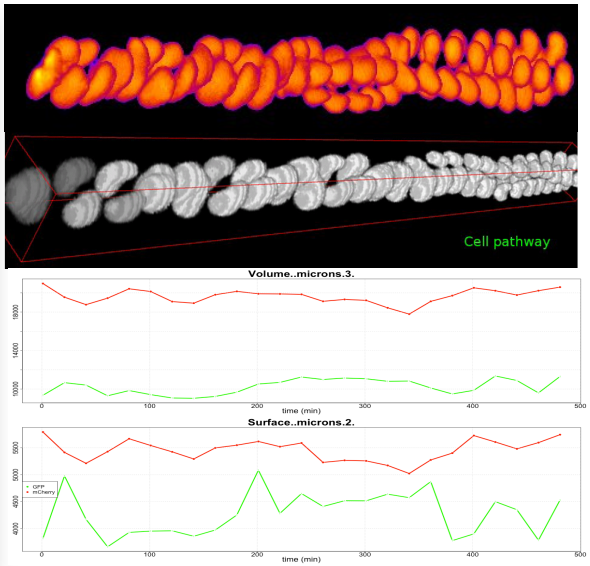
Recurrence

Source: Mark LaBarge

Cell identification and analysis







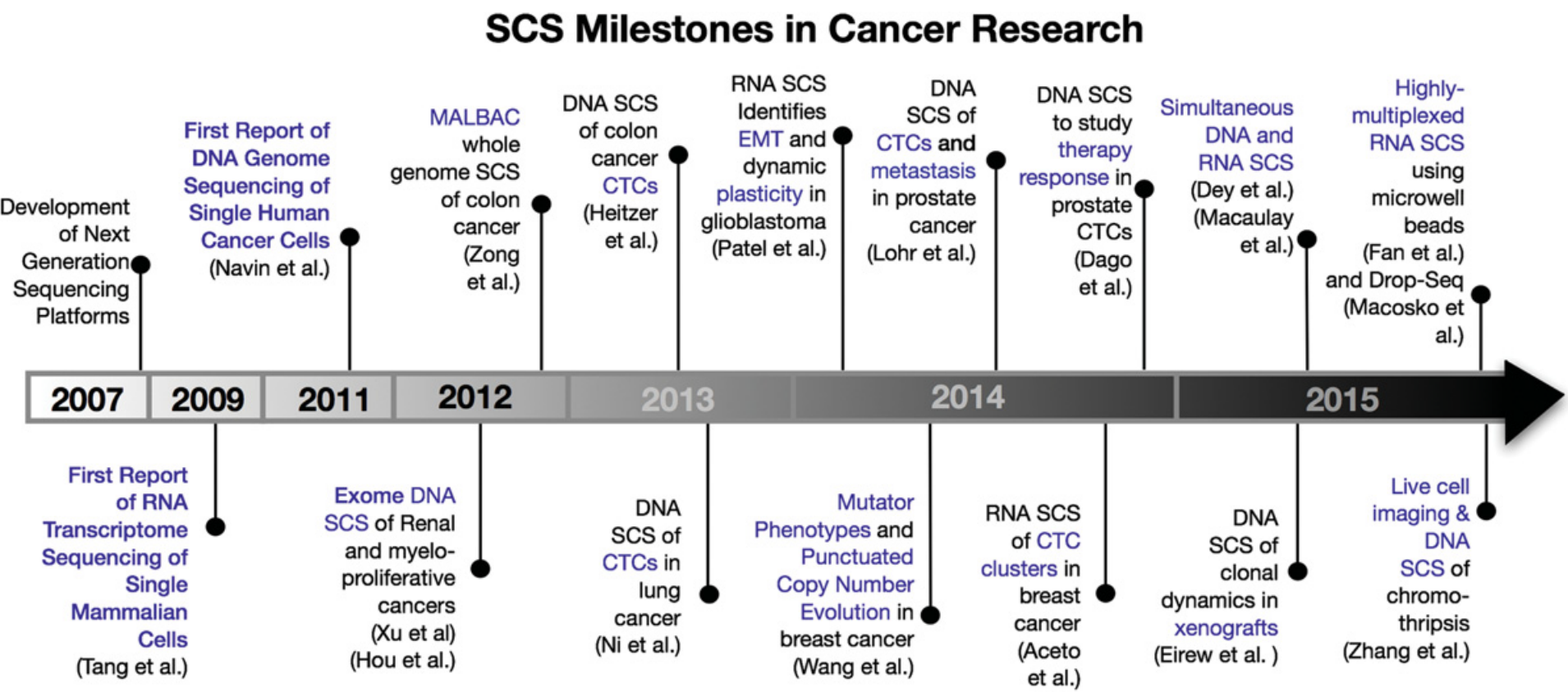
Fast method to analyze cervical cells: segment and identify subcellular components in 12 seconds

IEEE ISBE award

Quantitative time-lapse image analysis: confocal microscopy on breast cancer pathways in HMEC

Motion analysis: Associating motion as part of the tissue formation and final morphology

Combining Genome Analysis at Single Cell Level



The first five years of single-cell cancer genomics and beyond
Nicholas E. Navin, U. Texas

Collaboration models and example results

- **Institutes of expertise dedicated to Science of X**
  - CAMERA and Superfacility model
- **Access to HPC systems and performance expertise**
  - HipMer
- **Long-term software and data infrastructure**
  - KBase
- **Co-developing instruments and analysis tools**
  - Brain and CryoEM
- **Grand challenges**
  - Antibiotics
  - Cancer
  - Brain