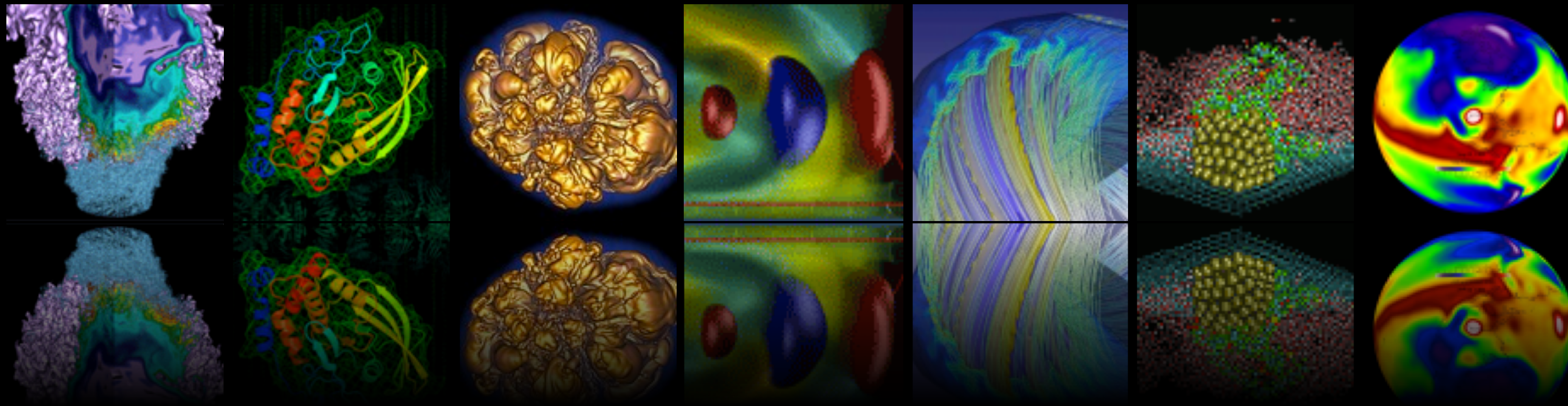


A Superfacility Model for Science

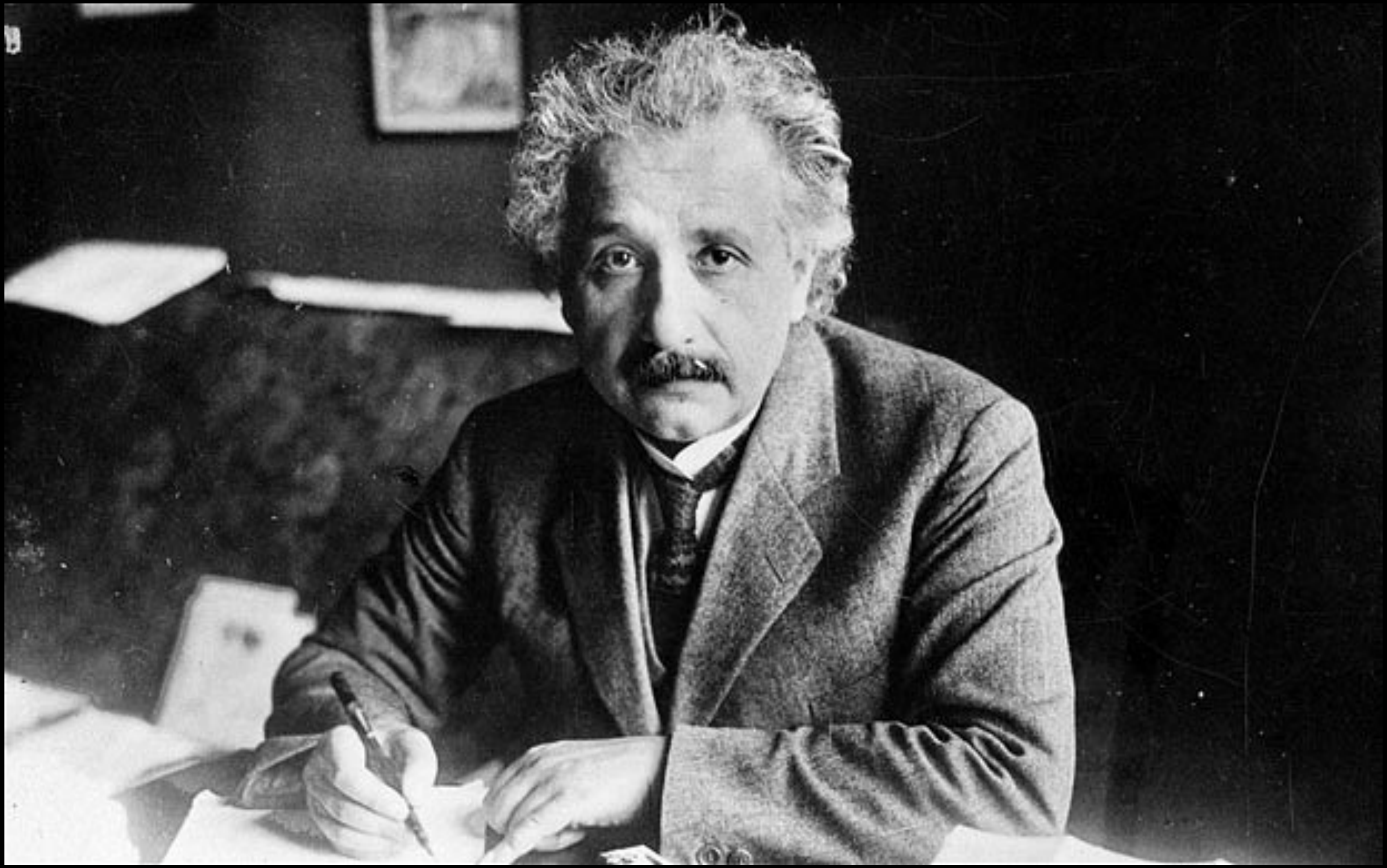
Kathy Yelick

Associate Laboratory Director for Computing Sciences
Lawrence Berkeley National Laboratory
Professor of Electrical Engineering and Computer Sciences
University of California at Berkeley

Part 1

Science is poised for transformation

Old School Scientists: The Lone Scientist



Team Science

New Scientists

17-year-old Brittany Wegner creates breast cancer detection tool that is 99% accurate on a minimally invasive, previously inaccurate test.

Machine Learning + Online Data + Cloud Computing

Experimental Science is Changing

By using our website you agree to our use of cookies in accordance with our cookie policy.

OK

PRIVACY POLICY +

Search for Mice

Advanced Mice Search

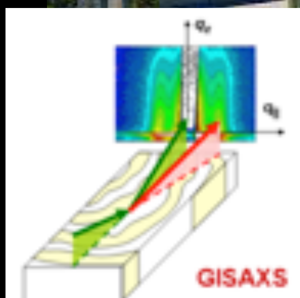
JAX® MICE & SERVICES

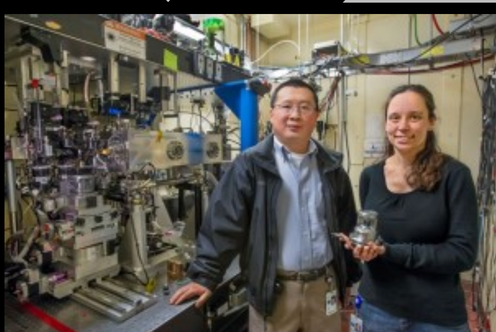
Search for mice by strain, stock, gene, allele and synonyms

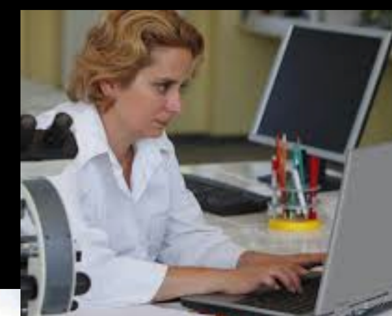
JAX® Mice are the highest quality and most-published mouse models in the world. Take advantage of our large inventories of common inbred strains and the convenience of having your breeding and drug efficacy needs met by the leading experts in mouse modeling.

- Breed Your Mouse
- Test Your Drug
- Cryopreserve Your Mouse

Old School Scientific Workflow



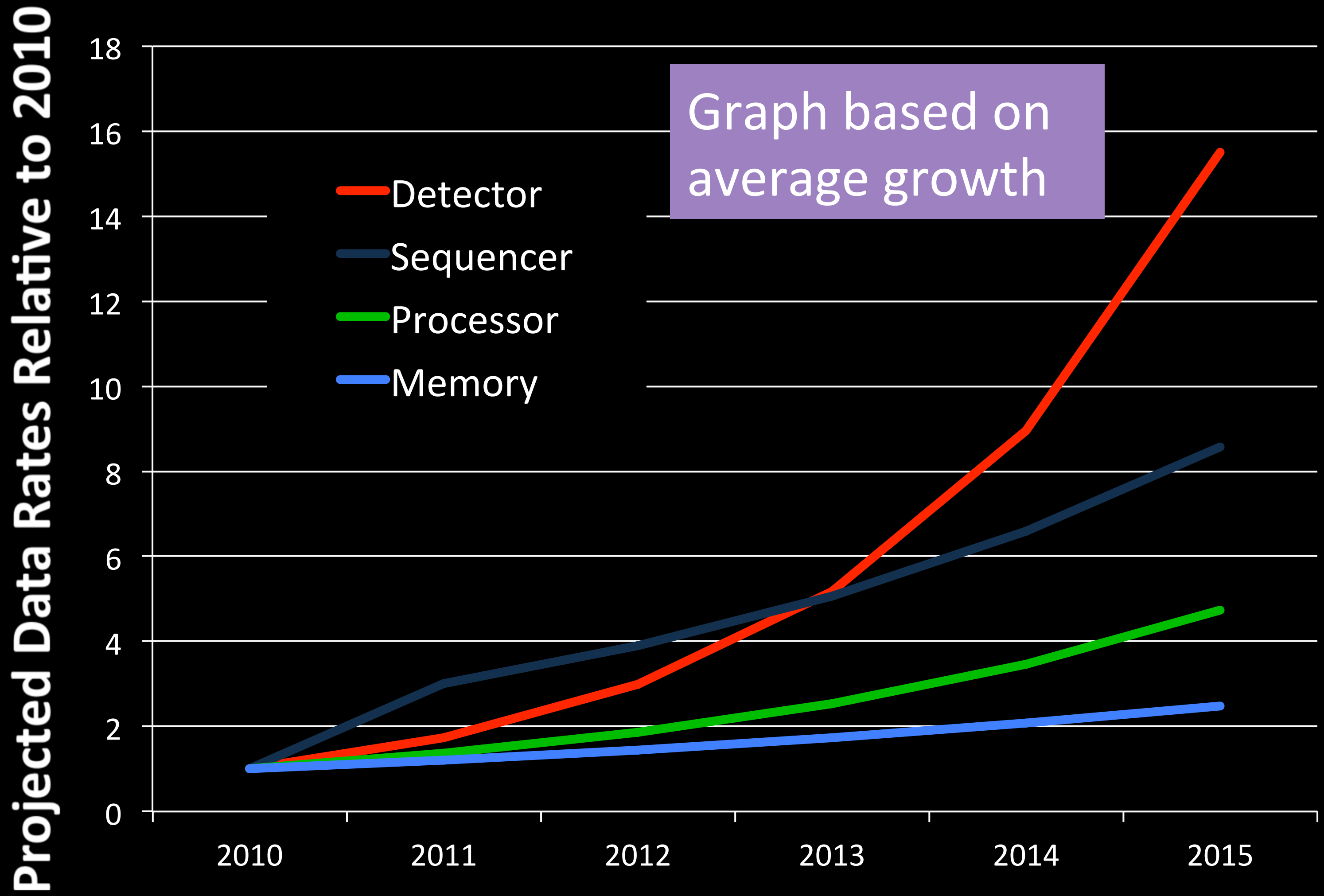




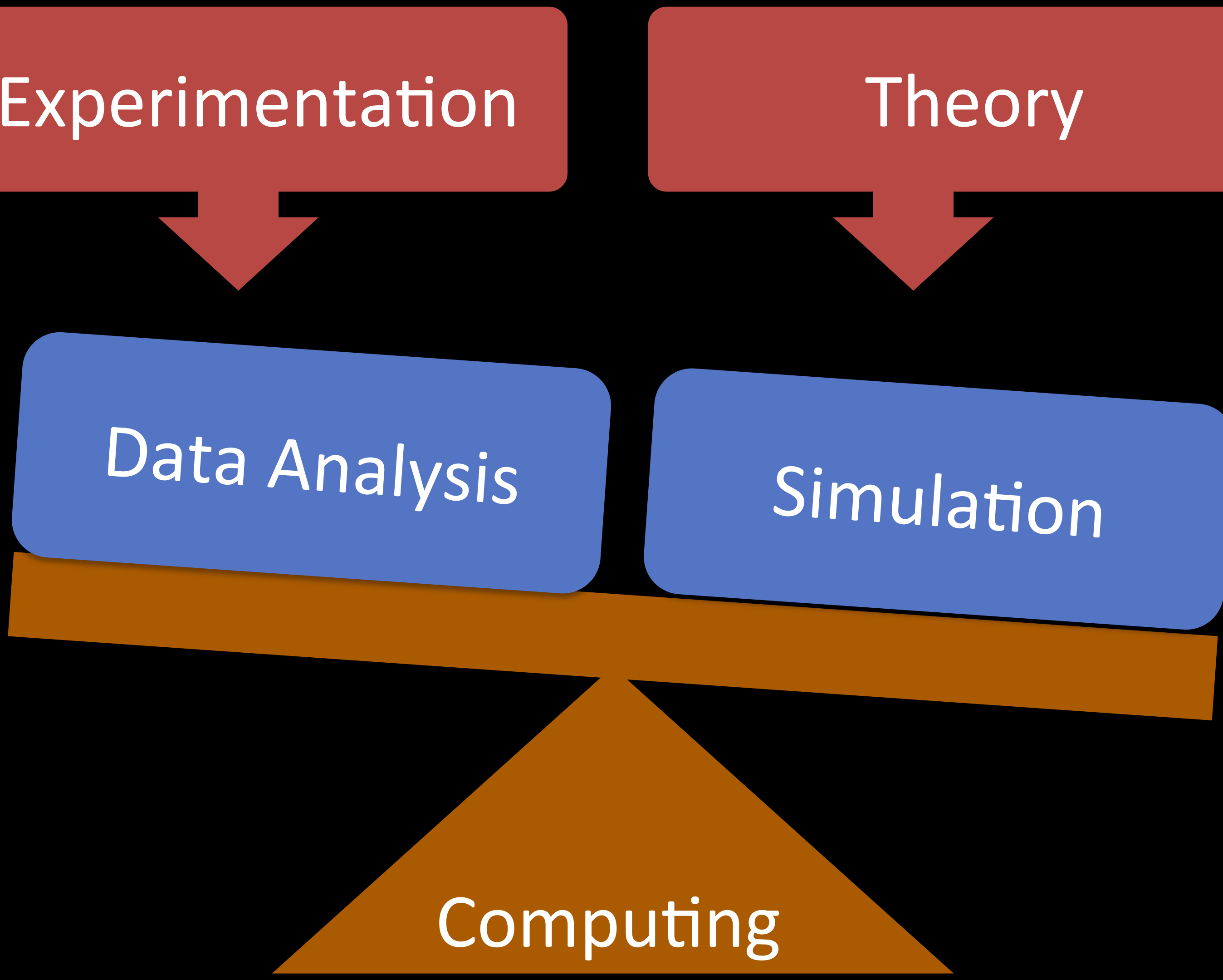
Computing, experiments, networking and expertise in a "Superfacility" for Science

Liu et al, "Fast printing and in situ morphology …". Adv Mater. 2015

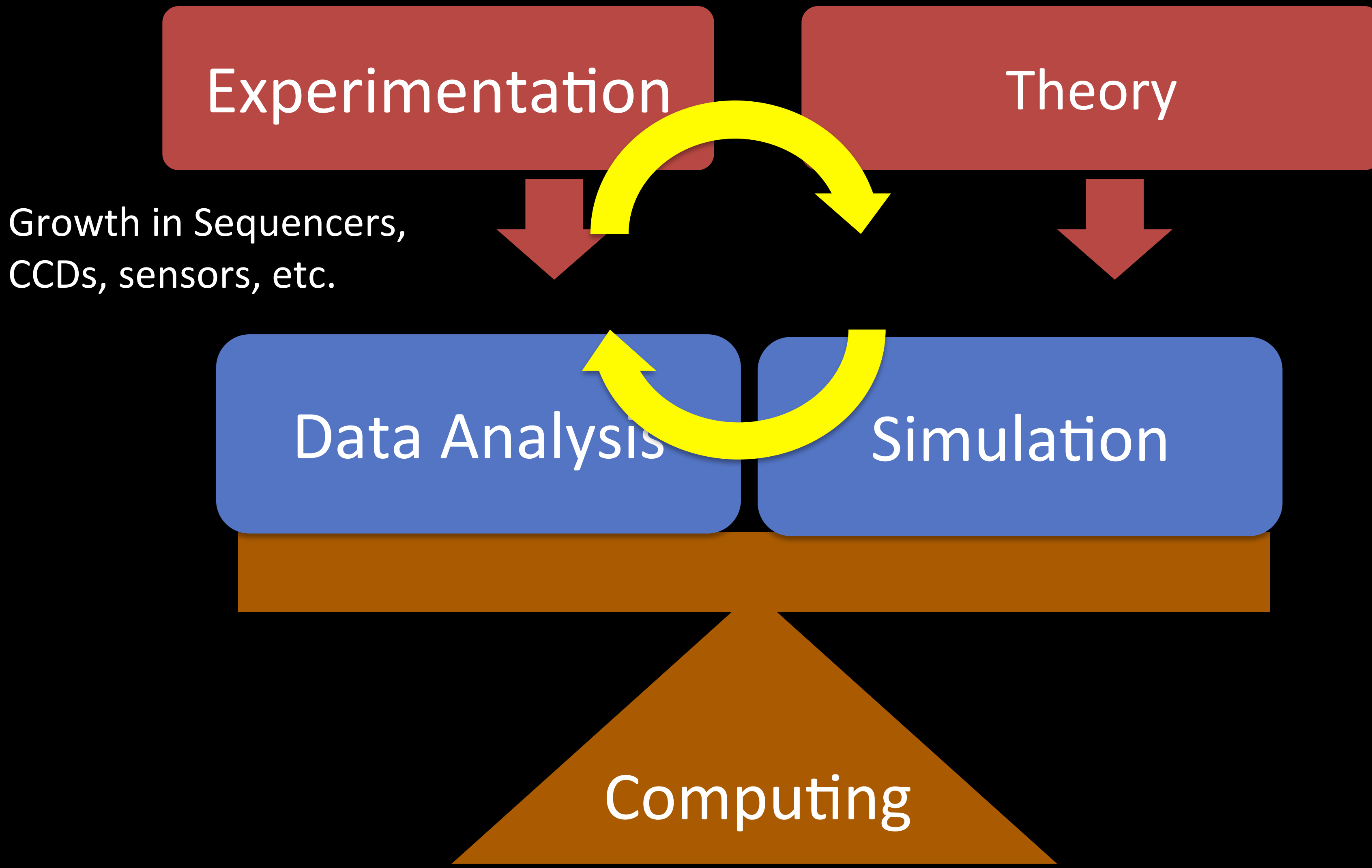
Data Growth is Outpacing Computing Growth



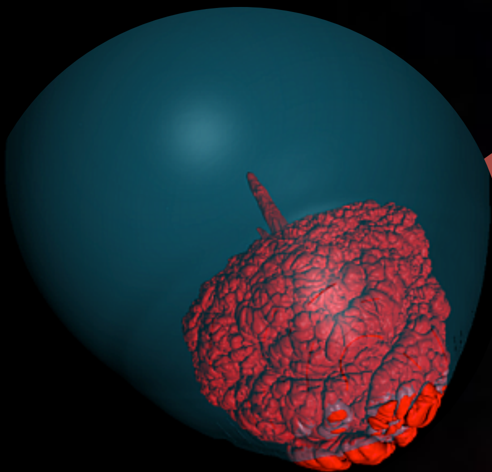
HPC: It's not just for simulation



HPC: It's not just for simulation

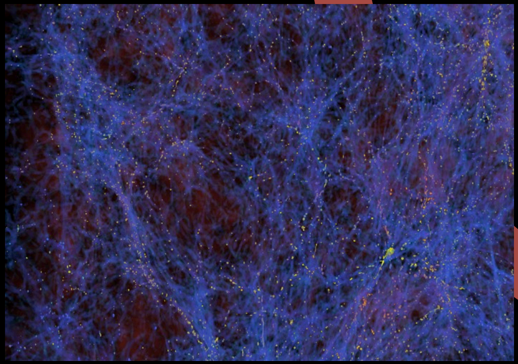


Integration of Simulation and Observational Science



A. Goobar, P. Nugent, et al (2017) Science

Intermediate Palomar Transient Factory with DESI, CMB-S4 and LSST coming



Simulations aid in interpreting data

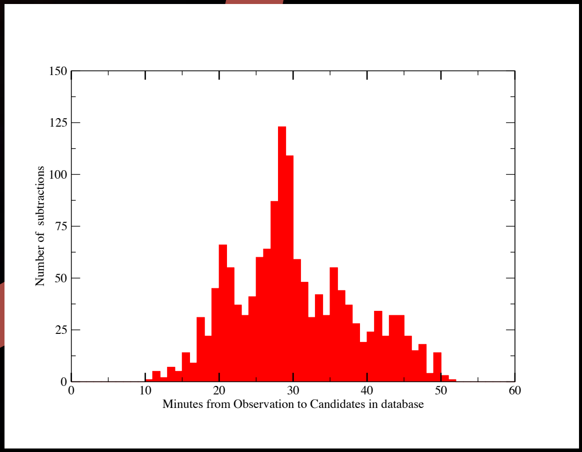
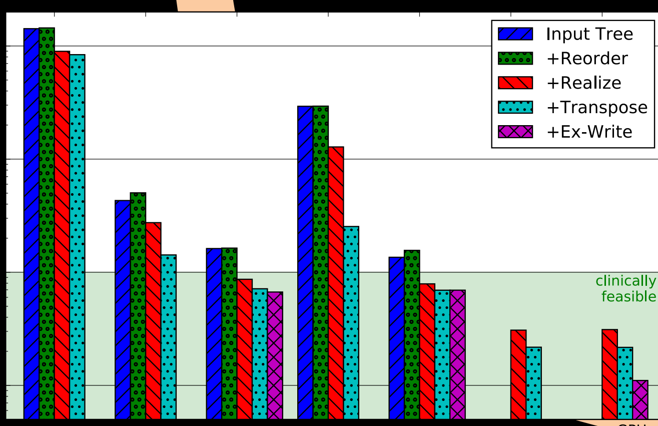


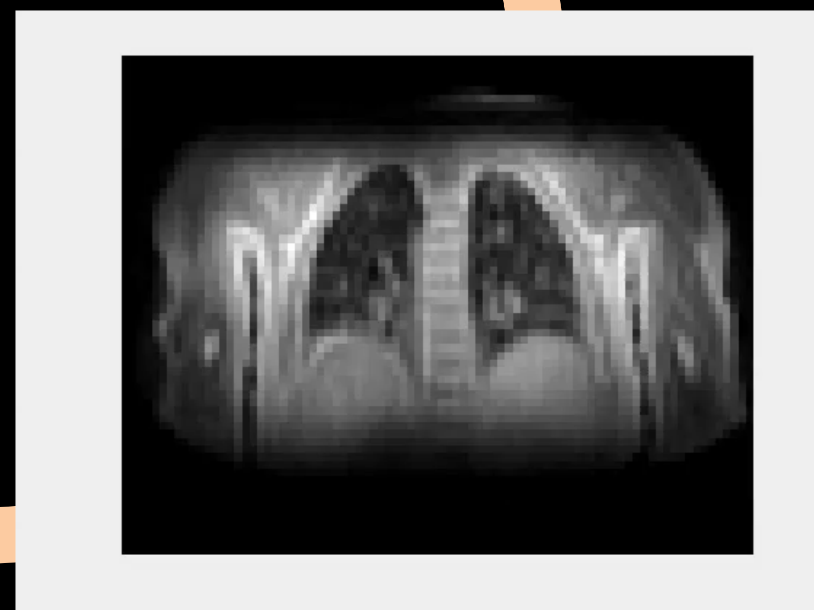
Image subtraction, machine learning in minutes

Real-Time MRI Challenge



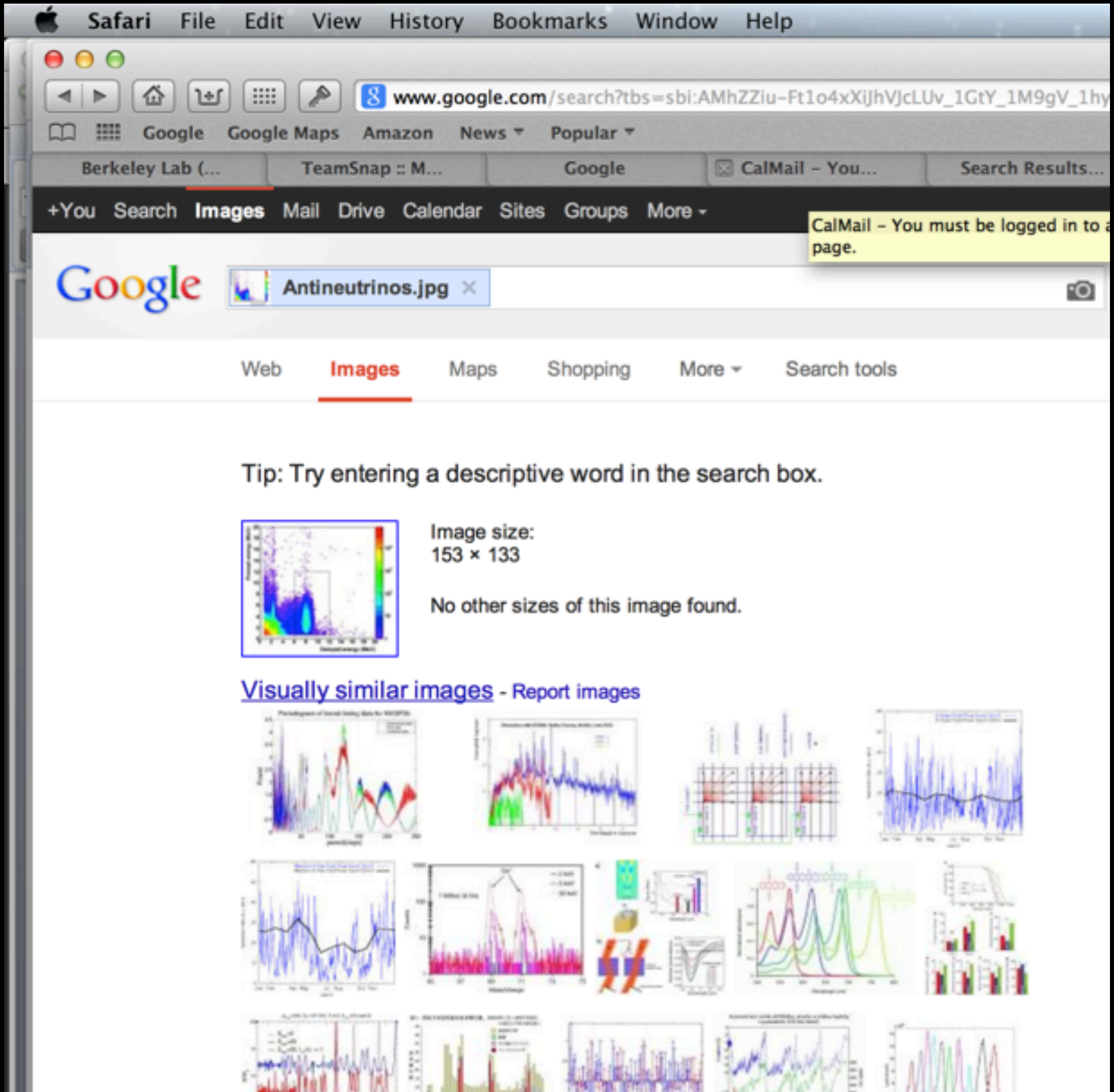
3 min goal (1 sec/iteration)

Michael Driscoll HPC optimization



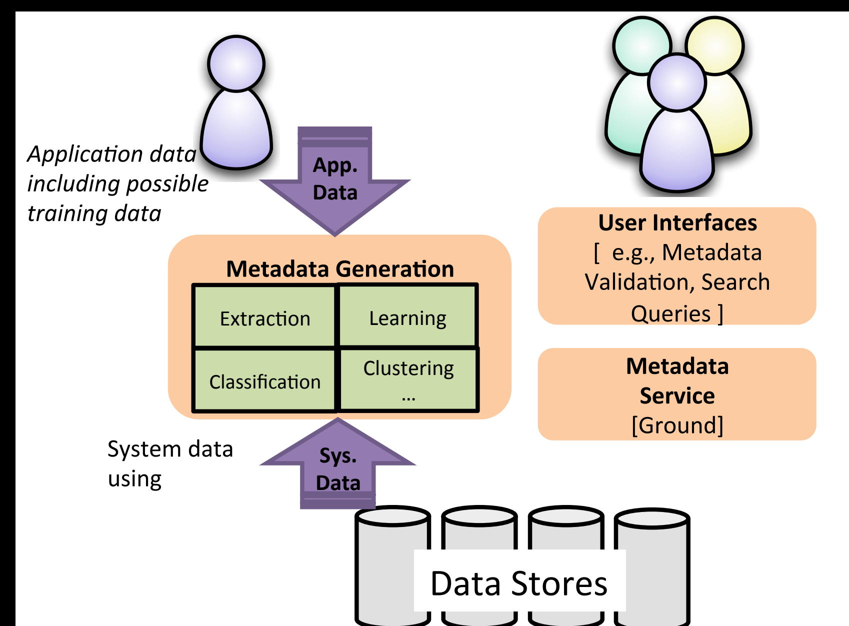
Compressed Sensing Approach by Mike Lustig et al
MRI results Wenwen Jiang

Old School Scientific Data Search



Automated Search, Meta-Data Analysis, and On-Demand Simulation

Automated metadata extraction using machine learning

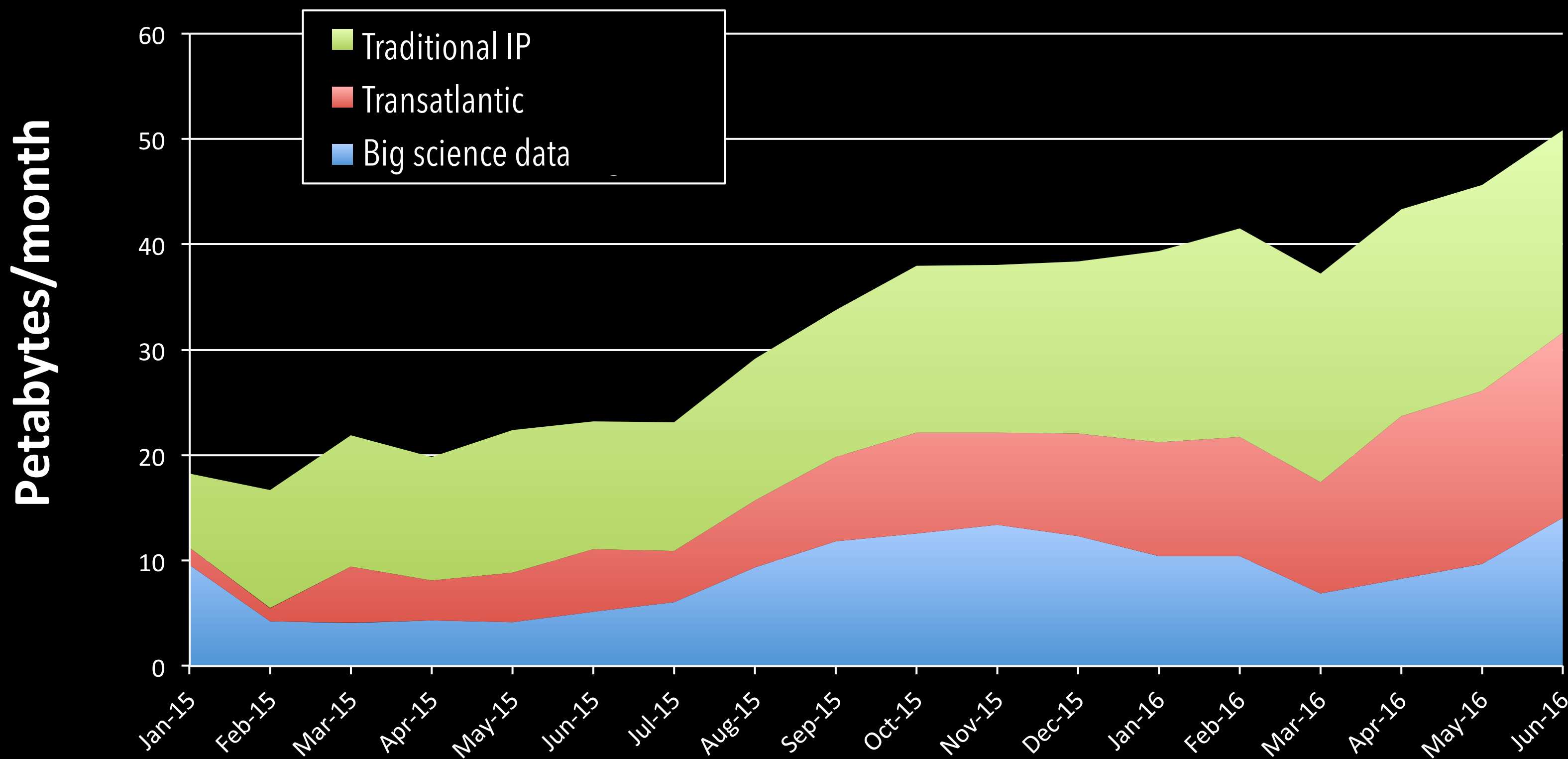


Jobs submitted by “bots” based on queries; algorithms extract informatics for design

Part 2

Computing and Networking Facilities need to adapt

ESnet: Exponential data growth drives capacity

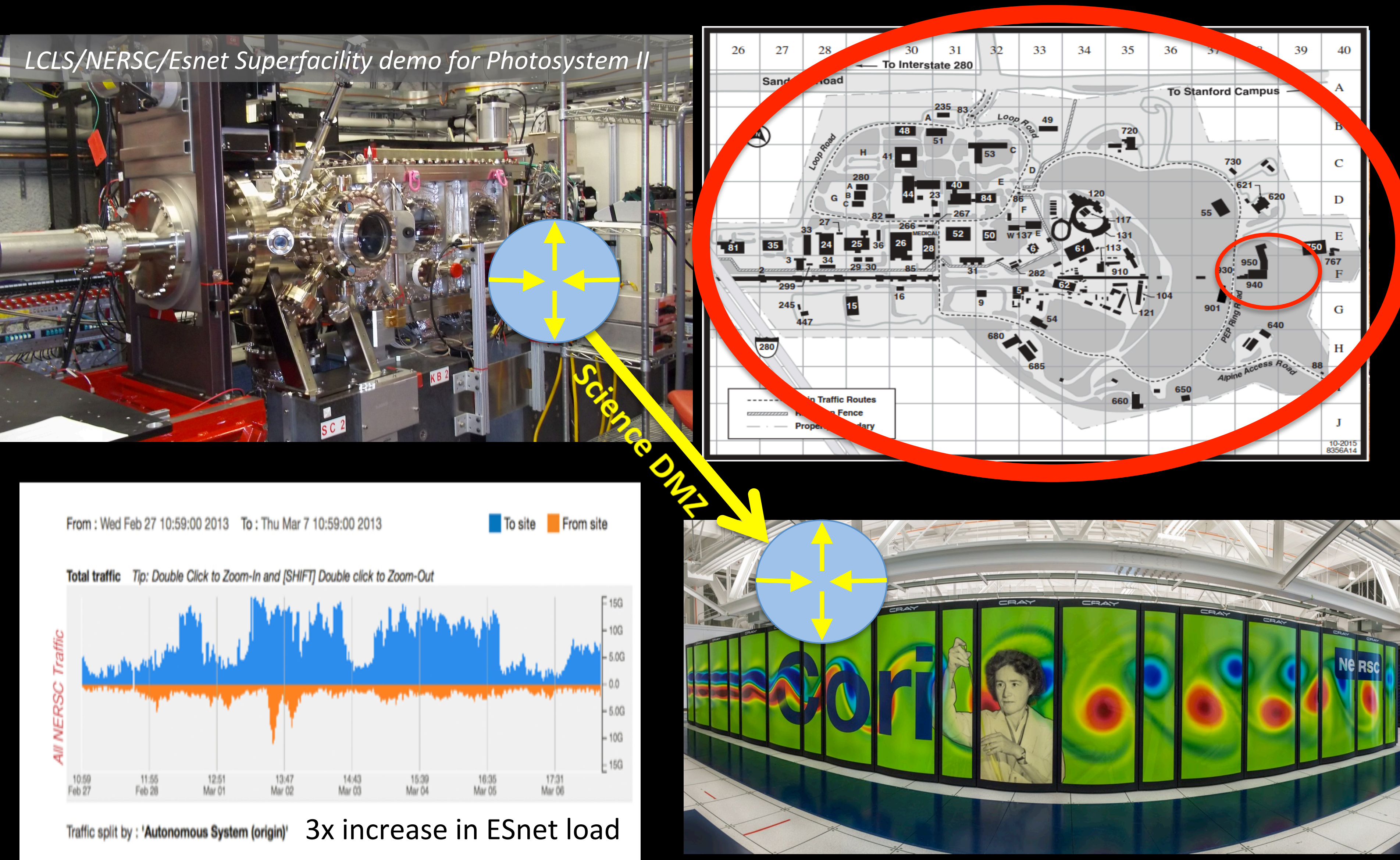


Science DMZ to deliver bandwidth to the end users

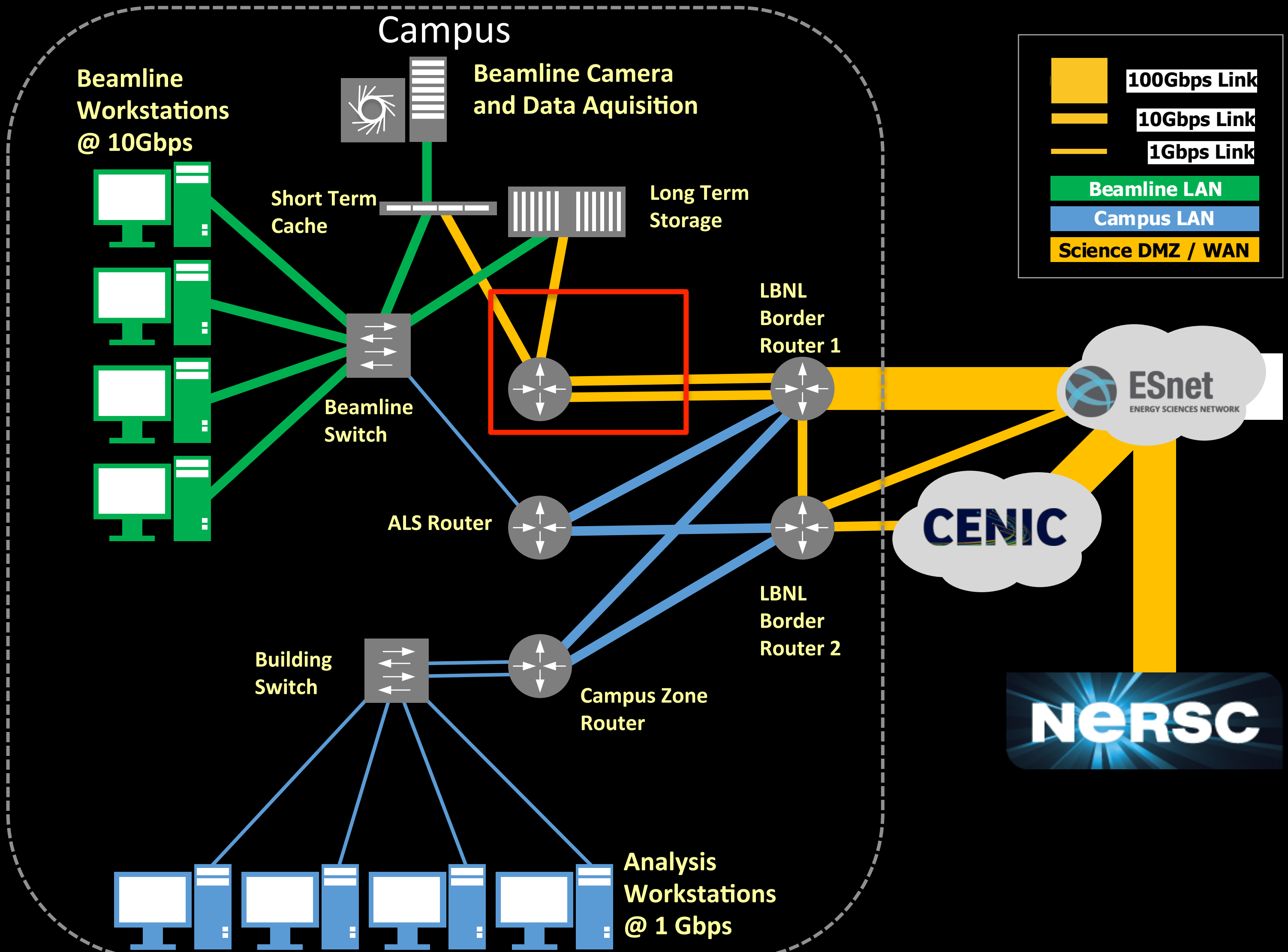
OSCARS for bandwidth reservation

100 Exabytes/year by 2024!

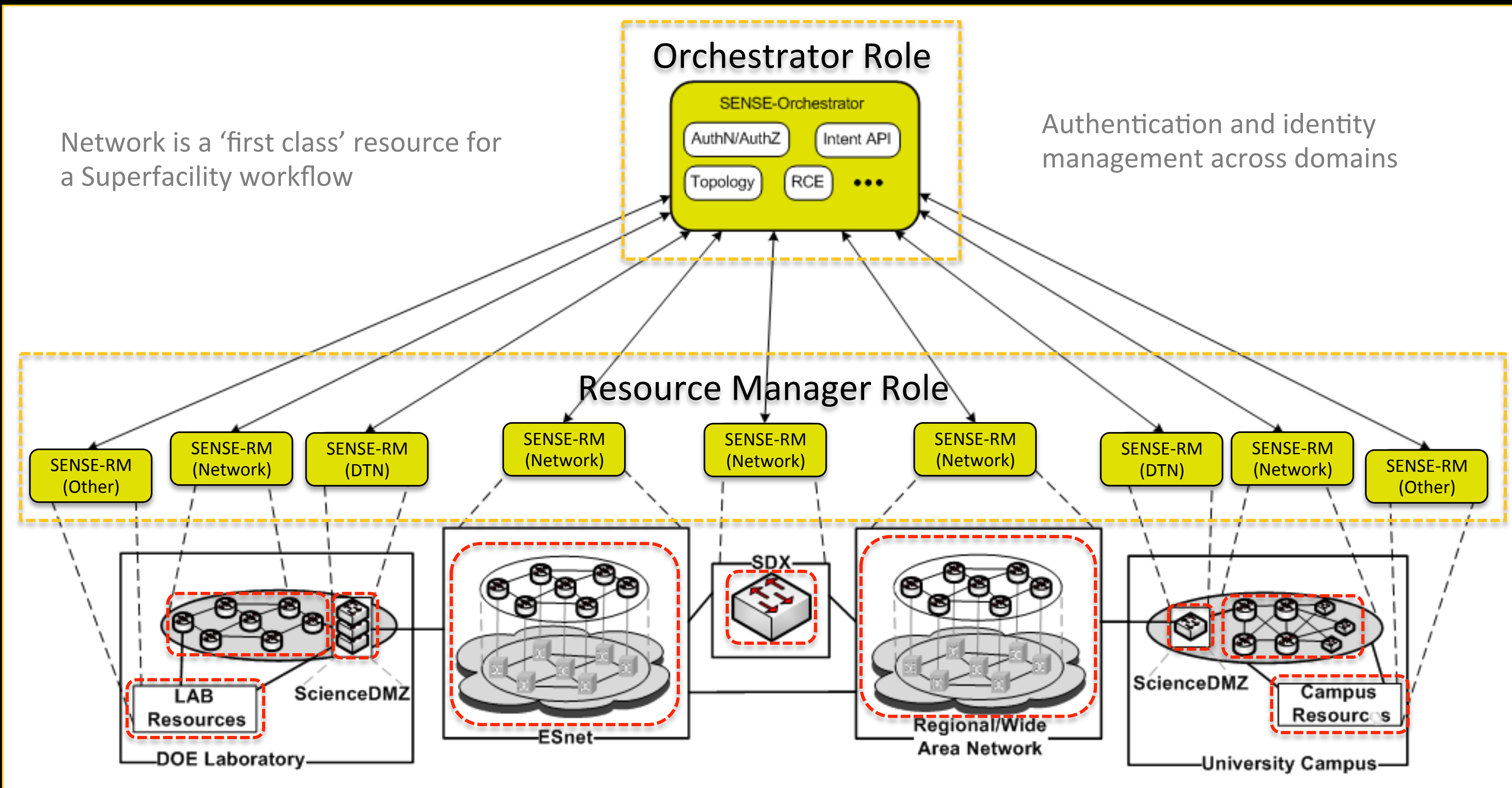
Bringing the Computer to the Experiment



Instruments and facilities require high-speed data network architectures like ScienceDMZ



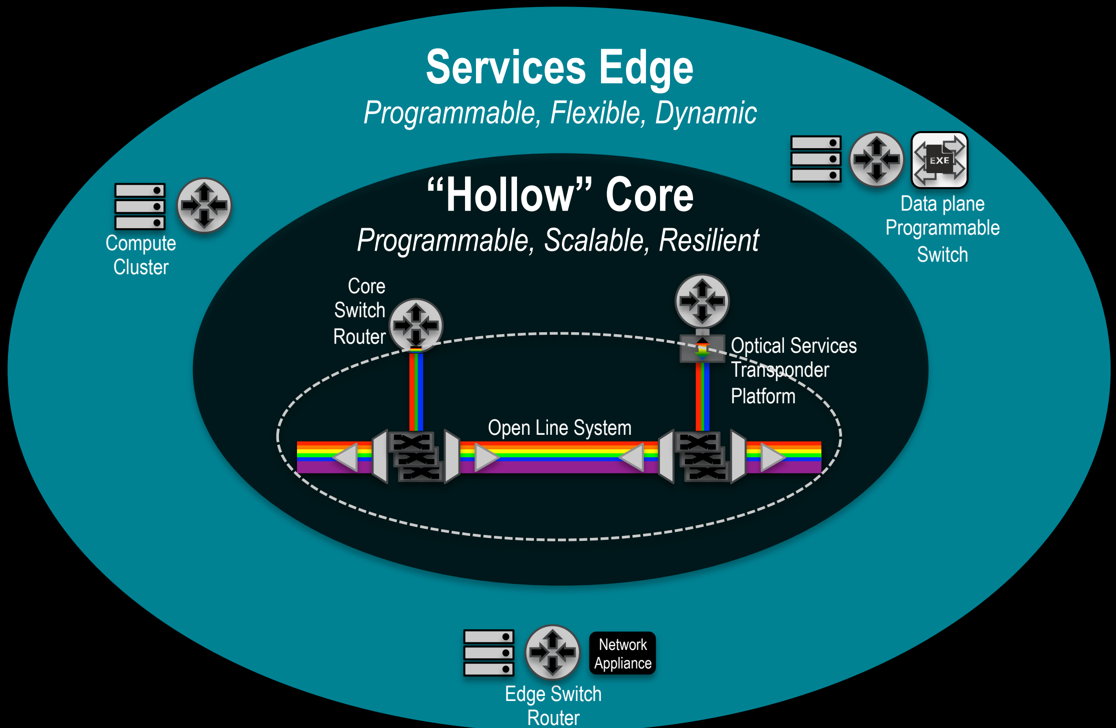
End-to-end, multi-domain network orchestration



Defines Service Perimeter/Boundary

SDN for end-to-end Network @ Exascale (**SENSE**) project led by Monga @ESnet with ANL, Caltech, FNAL, NERSC, MAX/UMD

ESnet6 plans for superfacility support



"Hollow" Core optimized for performance

- ***Programmable*** to allocate bandwidth and monitor status
- ***Scalable*** –Leverage latest technology (e.g. FlexGrid spectral partitioning, tunable wave modulation)
- ***Resilient*** – Protection and restoration functions using next generation Traffic Engineering (TE) protocols

Services Edge optimized for flexibility

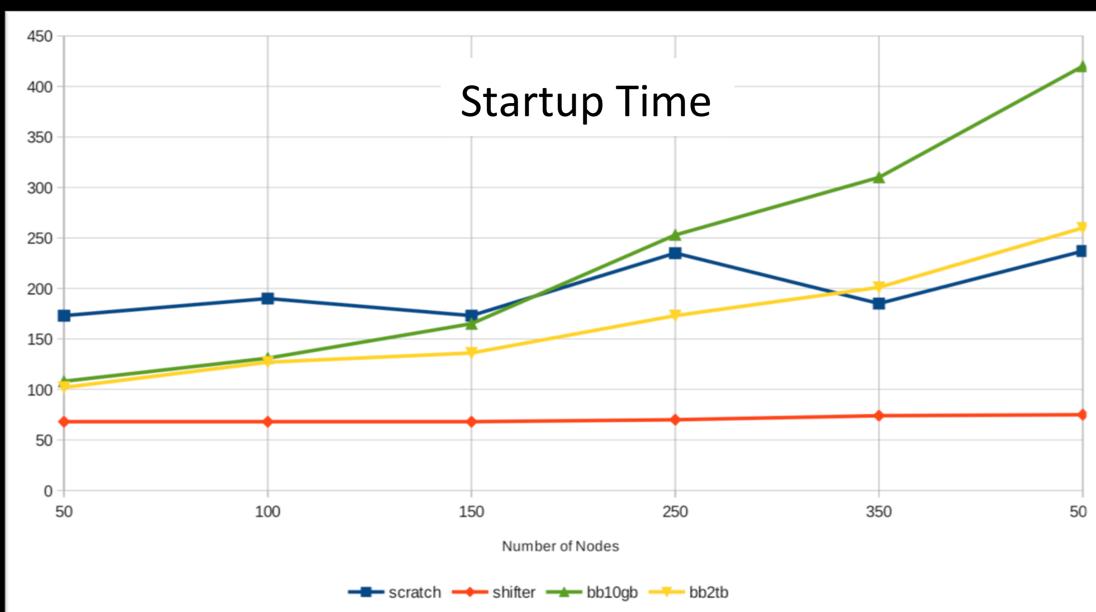
- ***Programmable*** to manage edge router/switch and retrieve telemetry information
- ***Flexible*** programmable switches (e.g. FPGA, NPU)
- ***Dynamic*** instantiation of services driven by SDN paradigms (e.g., virtualization, service chaining).

Systems configured for data-intensive science

NERSC Cori has data partition (Haswell) and pre-exascale (KNL)
NVRAM file system with close to 2 PB at 2 TB/sec
WAN-to-Cori optimized for streaming data: 100x faster from LCLS to Cori and Globus to CERN

Containers for HPC Systems

- Data analysis pipelines are often large, complex software stacks
- NERSC Shifter (with Cray), supports containers for HPC systems
- Used in HEP and NP projects
  (ATLAS, ALICE, STAR, LSST, DESI)




The Register®
Biting the hand that feeds IT
DATA CENTER SOFTWARE NETWORKS SECURITY INFRASTRUCTURE DEVOPS BUSINESS HARDWA

Data Center ▸ HPC

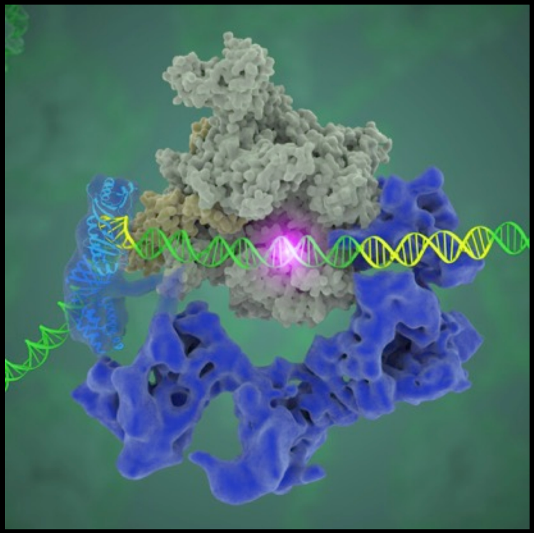
Cray hoists Docker containers into supercomputers

Productivity gains without performance hits

18 Nov 2015 at 00:01, Drew Cullen 25 35

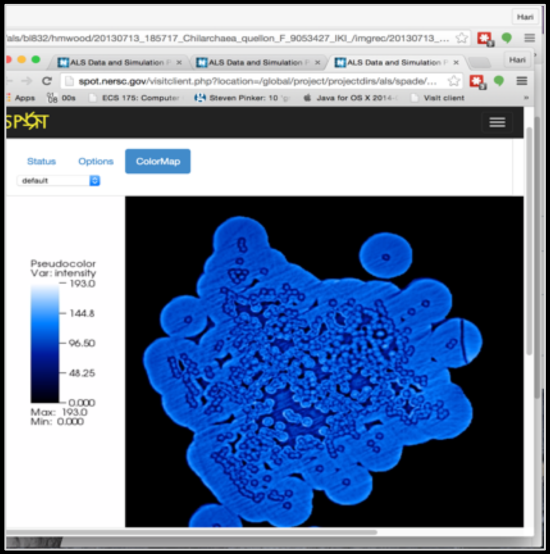
Real-time queue prototyped at NERSC

- **In 1998 dedicated hardware; now prototype queue on Cori**
- **<1% of NERSC allocation**
- **Cryo-Em, Mass spec, Telescopes, Accelerator, Light sources**



Cryo-EM: Image classification
Nogales Lab

PTF: Image subtraction pipeline

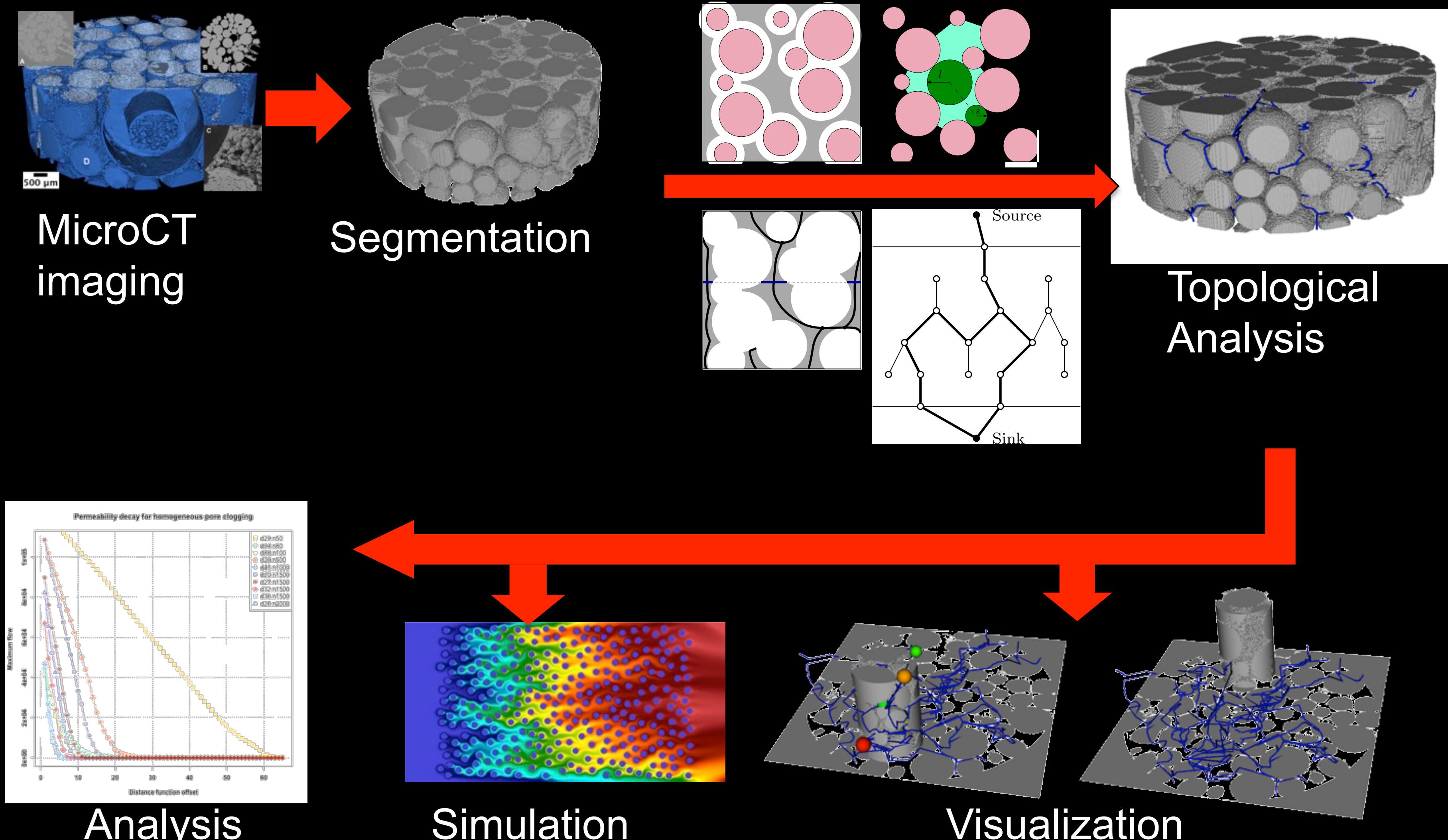


ALS: 3D Reconstruction,
rendered on SPOT web portal

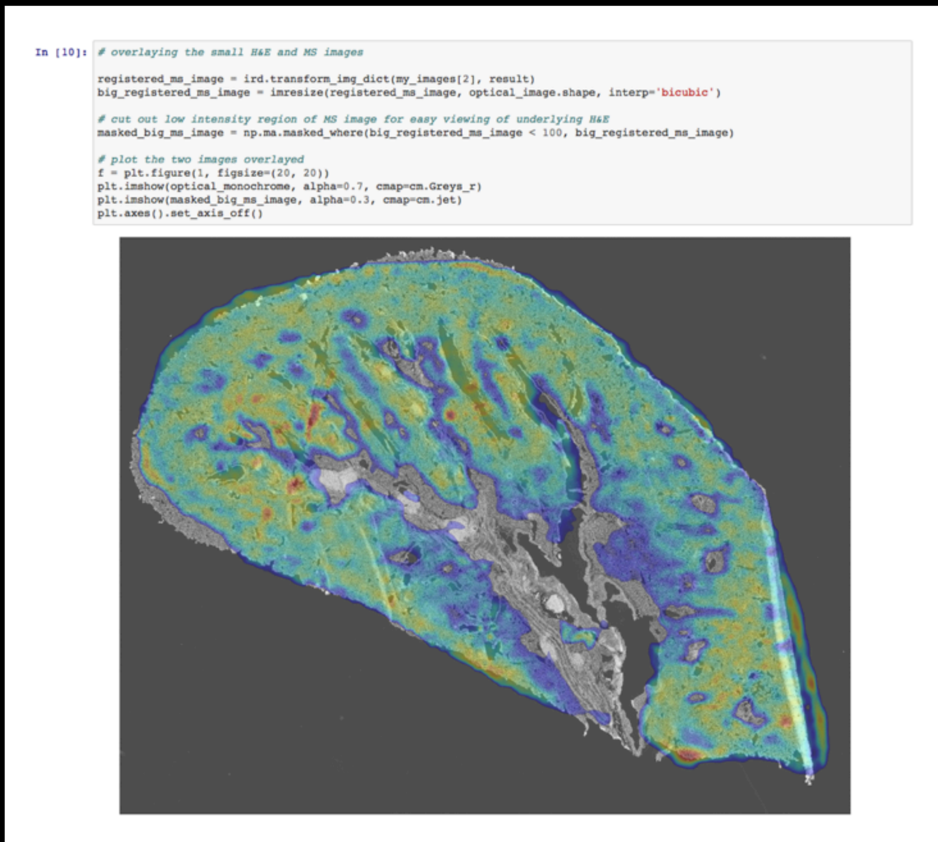
Part 3

Research challenges are substantial

Software implementations at scale in pipeline



Interactive Analytics using Jupyter



jupyter

Science notebooks through Jupyter (iPython)

- Widely used in science
- Interactive HPC LDRD

Deployed at NERSC:

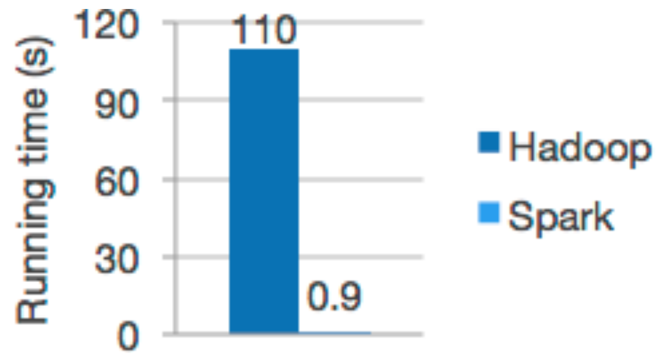
- >100 users pre-production

Fernando Perez et al

Productive Programming

Speed

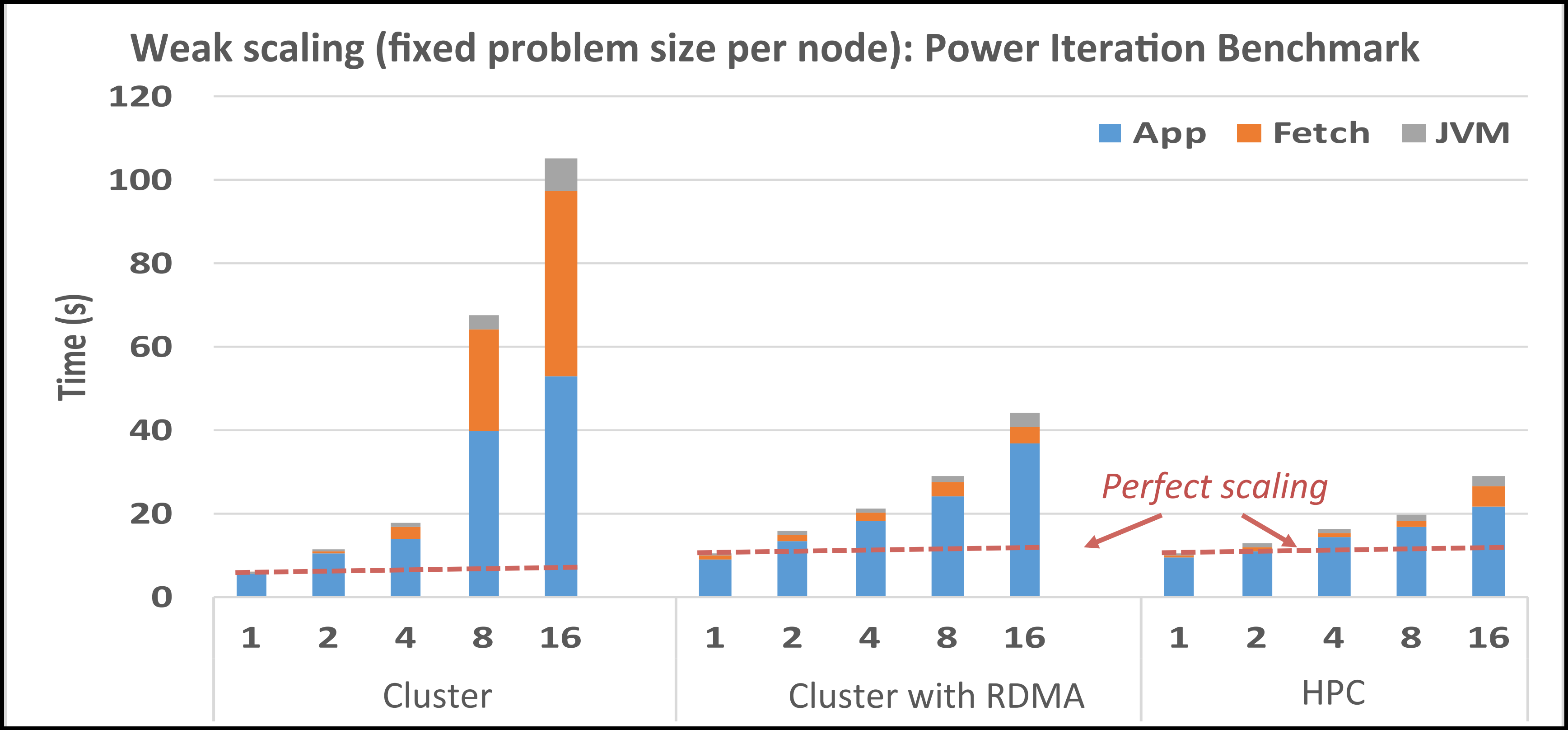
Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



- High failure rate
- Slow network
- Fast (local) disk

And Spark is still 10x+ slower than MPI

SPARK Analytics on HPC

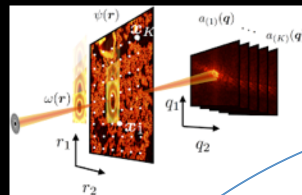


SPARK on HPC vs. clusters

- Network, I/O, and virtualization all key to performance
- Increased scale from O(100) to O(10,000) cores

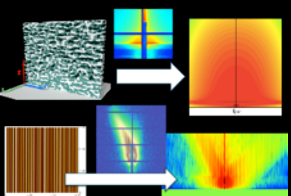
Chaimov, Malony, Iancu, Ibrahim, Canon, Srinivasan

CAMERA: Math for the Facilities



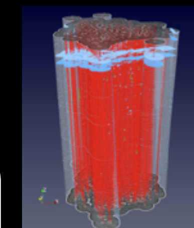
Designing mathematical algorithms to allow real-time analysis next to the equipment

Real-time streaming ptychography—ALS, delivered to NSLS2, LANL, BESSY,



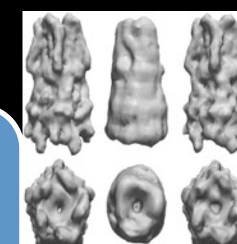
Multi-modal: Building the math that fuses information from multiple experiments

SFM/TEM + GISAXS



New algorithms to transform manual into automatic analysis

Automatic image processing for the ALS/GE

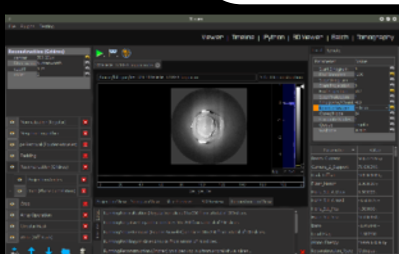


Inventing new math and models to match new acquisition technologies

Fluctuation scattering and single particle imaging for the LCLS

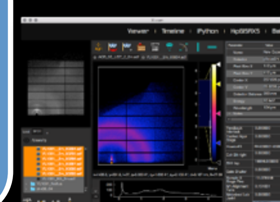
Compare and integrate multiple analysis tools

CAMERA workshop on Tomography: Joint with APS, ESRF, DIAMOND, LNLS, LLNL, SSRL,….,



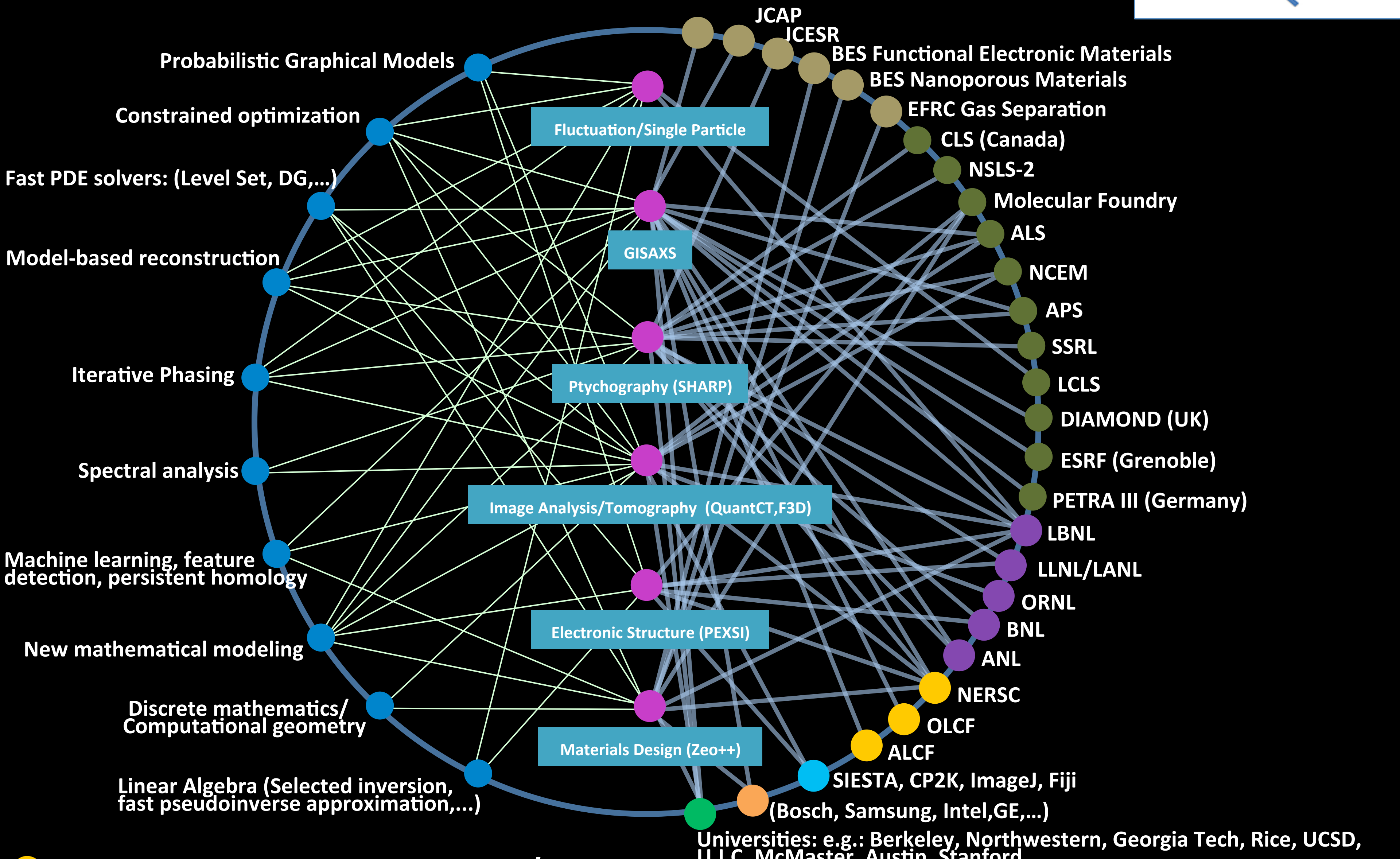
Cultural and Sociological Challenges

Robust and reliable codes and data flow: workflow environments



Workflow and access to remote supercomputers: XiCAM for ALS, SSRL, APS, NSLS2

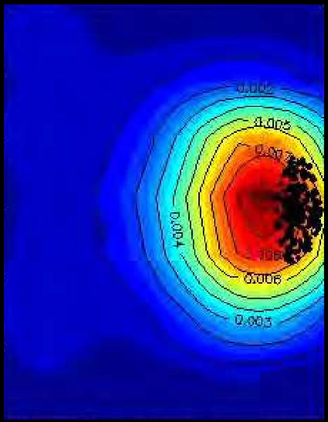
CAMERA: Mathematics for Facilities



=ASCR Facilities =BES Centers/Internat.

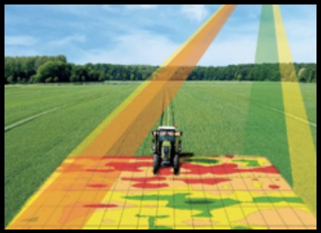
=BES (and other) Facilities =LABS =3[Rd] Party Codes =Industry =Universities

Machine Learning for Science

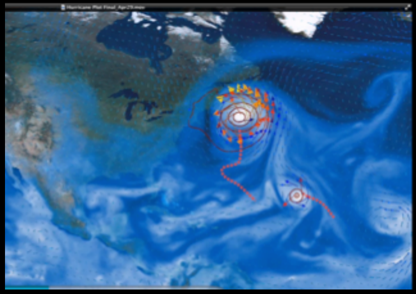


Accelerators

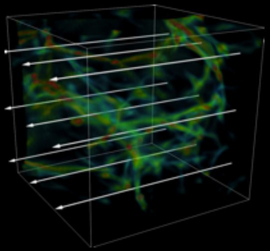
Images in cosmology, light sources, etc.



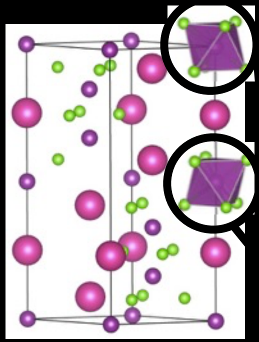
Biology



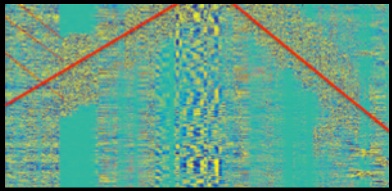
Climate

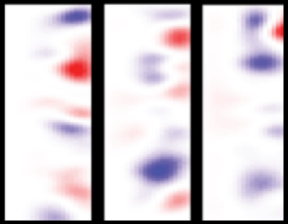


Cosmology simulation



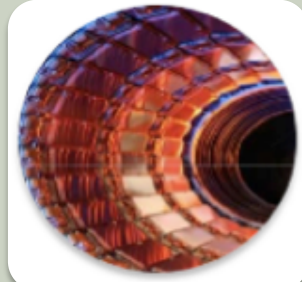
Chemistry





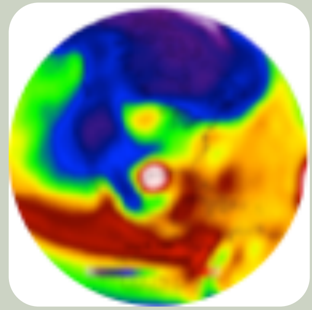
Data Complexity → Interpretability → Performance and Scale

Architectures for Data vs. Simulation

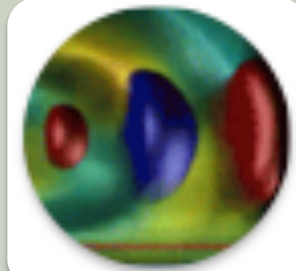


Separate Jobs

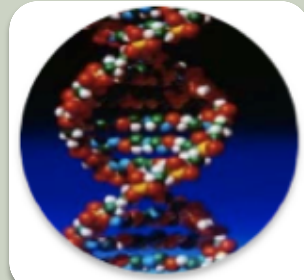
Compute Intensive



Nearest Neighbor



All-to-All



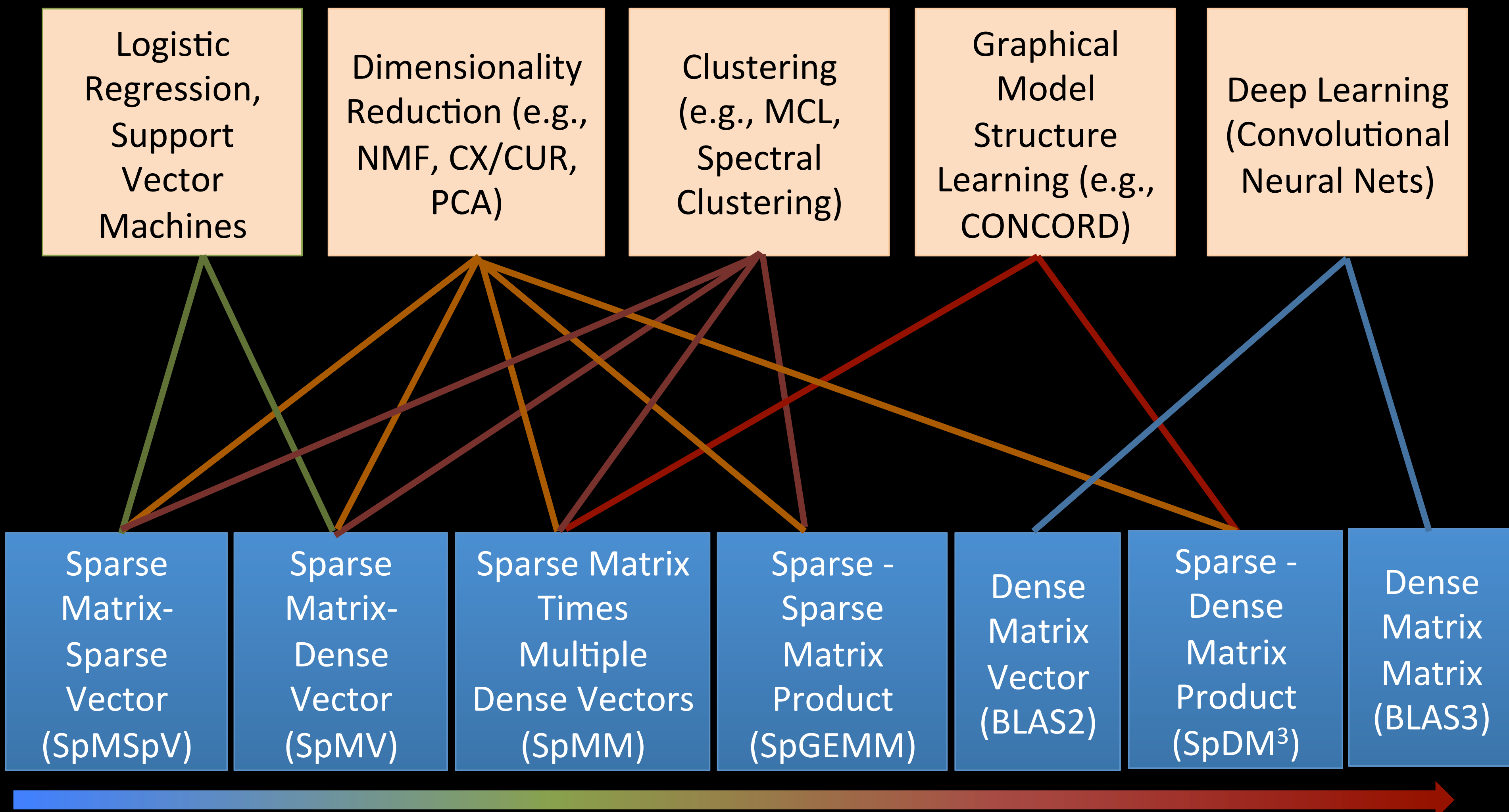
Random Access

Different architectures for simulation? Can simulation use data architectures?

Analytics vs. Simulation Kernels:

7 Giants of Data	7 Dwarfs of Simulation
Basic statistics	Monte Carlo methods
Generalized N-Body	Particle methods
Graph-theory	Unstructured meshes
Linear algebra	Dense Linear Algebra
Optimizations	Sparse Linear Algebra
Integrations	Spectral methods
Alignment	Structured Meshes

Machine Learning Mapping to Linear Algebra



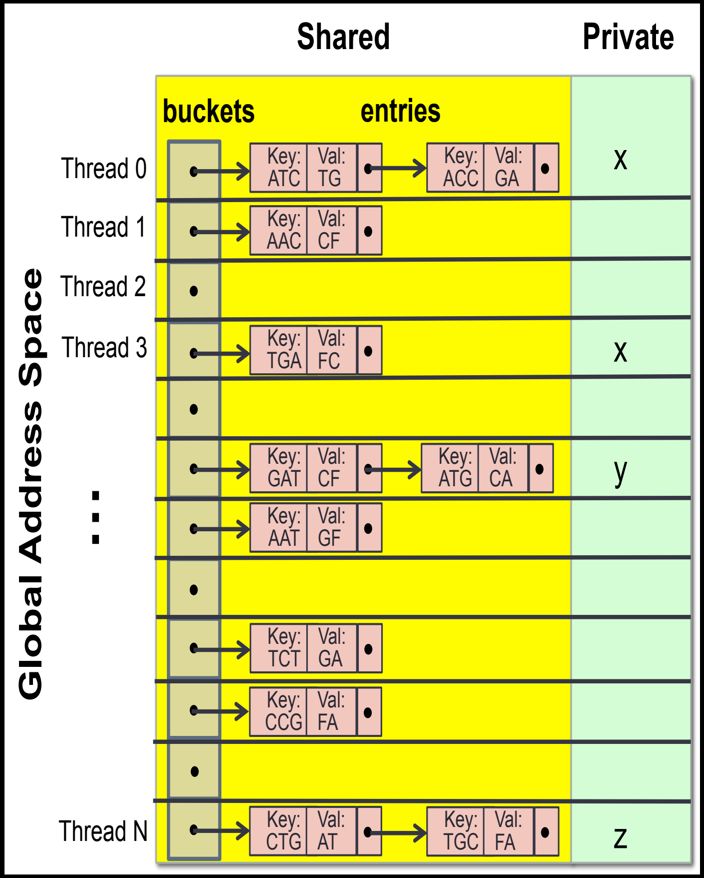
Aydin Buluc

Random Access Analytics

- Genome assembly “~~needs shared memory~~”

Global Address Space

- Low overhead communication
- Remote atomics
- Partitions for any structure



Scales to 15K+ cores

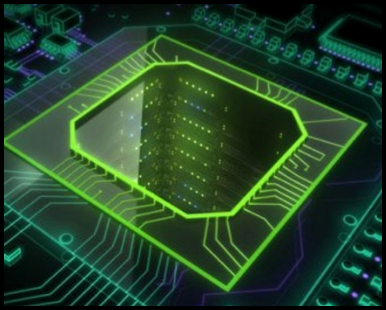
Under 10 minutes for human

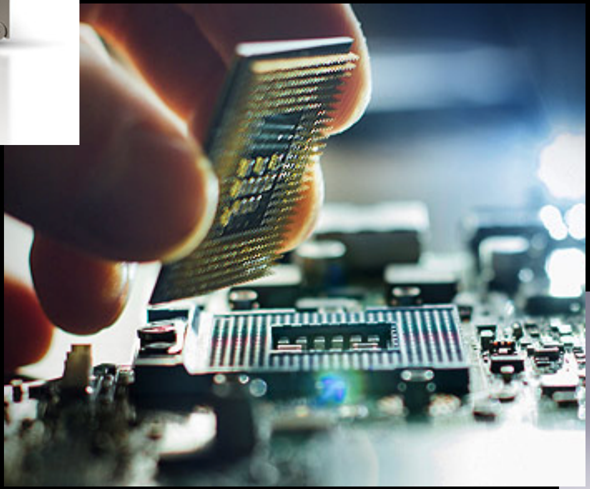
First ever solution

E. Georganas, A. Buluc, J. Chapman, S. Hofmeyr, C. Aluru, R. Egan, L. Oliker, D. Rokhsar, K. Yelick

Specialization: End Game for Moore's Law

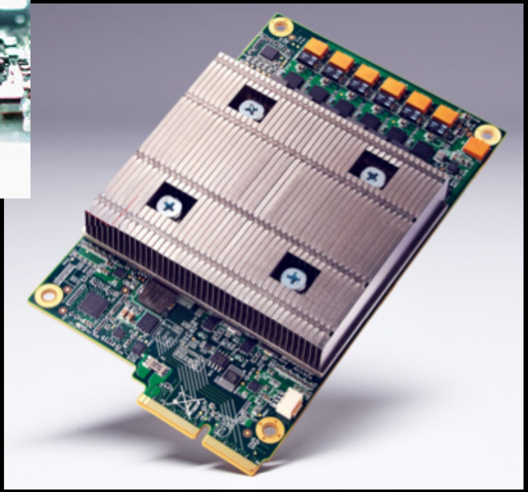
NVIDIA builds deep learning appliance with P100 Tesla's





Intel buys deep learning startup, Nervana

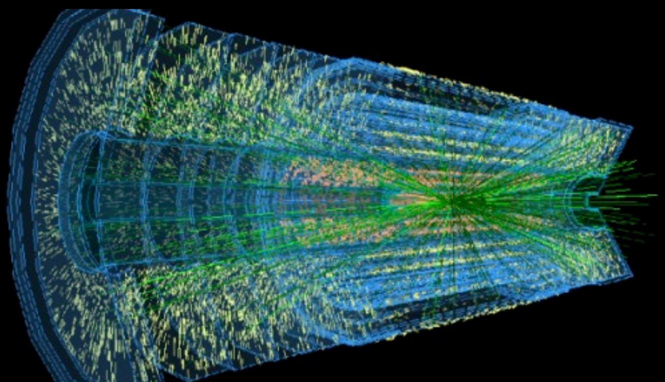
FPGAs

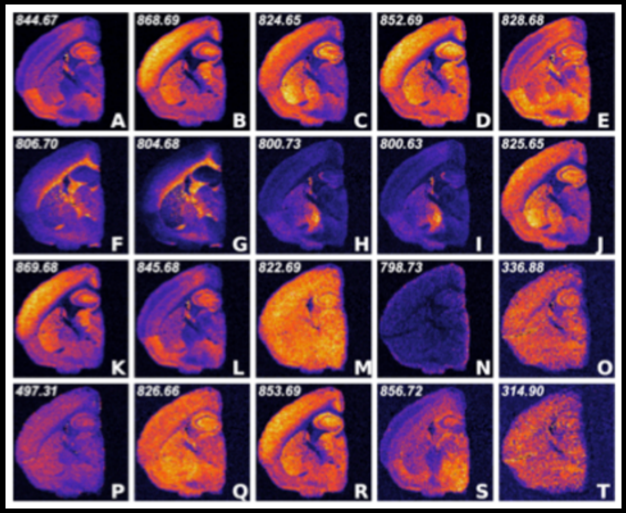


Google designs its own Tensor Processing Unit (TPU)

Data processing with special purpose hardware

- General trend towards specialization for performance
- Data processing (on raw data) will be first in DOE





Particle Tracking with Neuromorphic chips

Deep learning processors for image analysis

Computing in Detectors

FPGAS for genome analysis

Filtering, De-Noise and Compressing Data

AmeriFlux & FLUXNET: 750 users access carbon sensor data from 960 carbon flux data years

Arno Penzias and Robert Wilson discover Cosmic Microwave Background in 1965

Superfacility Vision

Terabit detectors

Detectors send terabit data streams.

Programmable Networks

Distributed facility provides resilience, accessibility

In situ analysis

Advanced algorithms and specialized hardware near detectors

Data platforms

High performance memory and storage systems

Automation and control

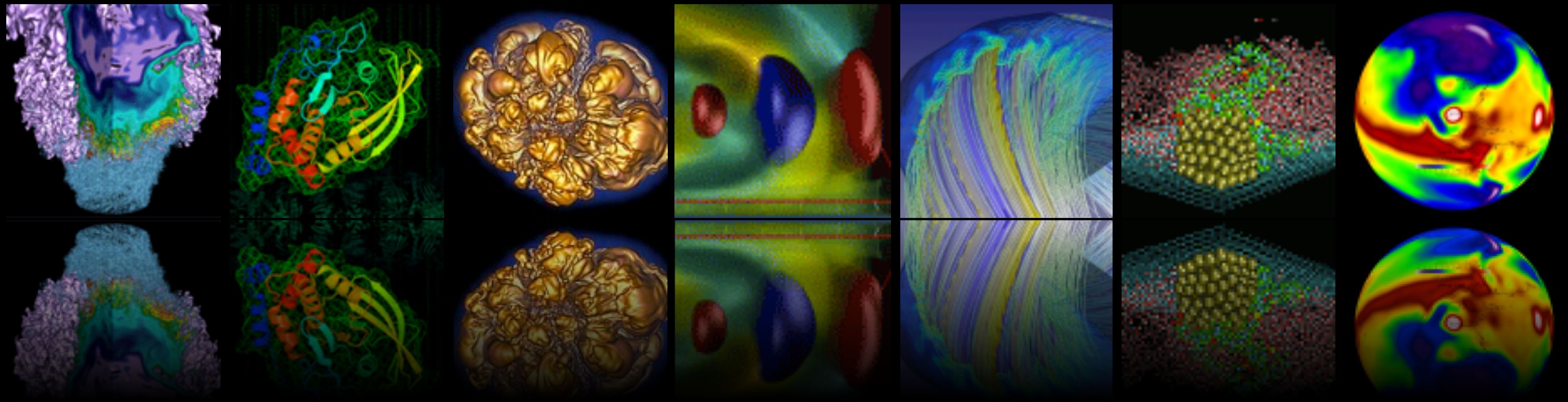
Real-time feedback to adjust telescopes or shot setup for experiments

Archive, sharing, and reuse

Data is archived and retrieval through automated metadata

Offline analysis

Combine with simulation to interpret data, improve models



Extreme Data Science

The scientific process is poised to undergo a radical transformation based on the ability to access, analyze, simulate and combine large and complex data sets.

Superfacility: Integrated network of experimental and computational facilities and expertise