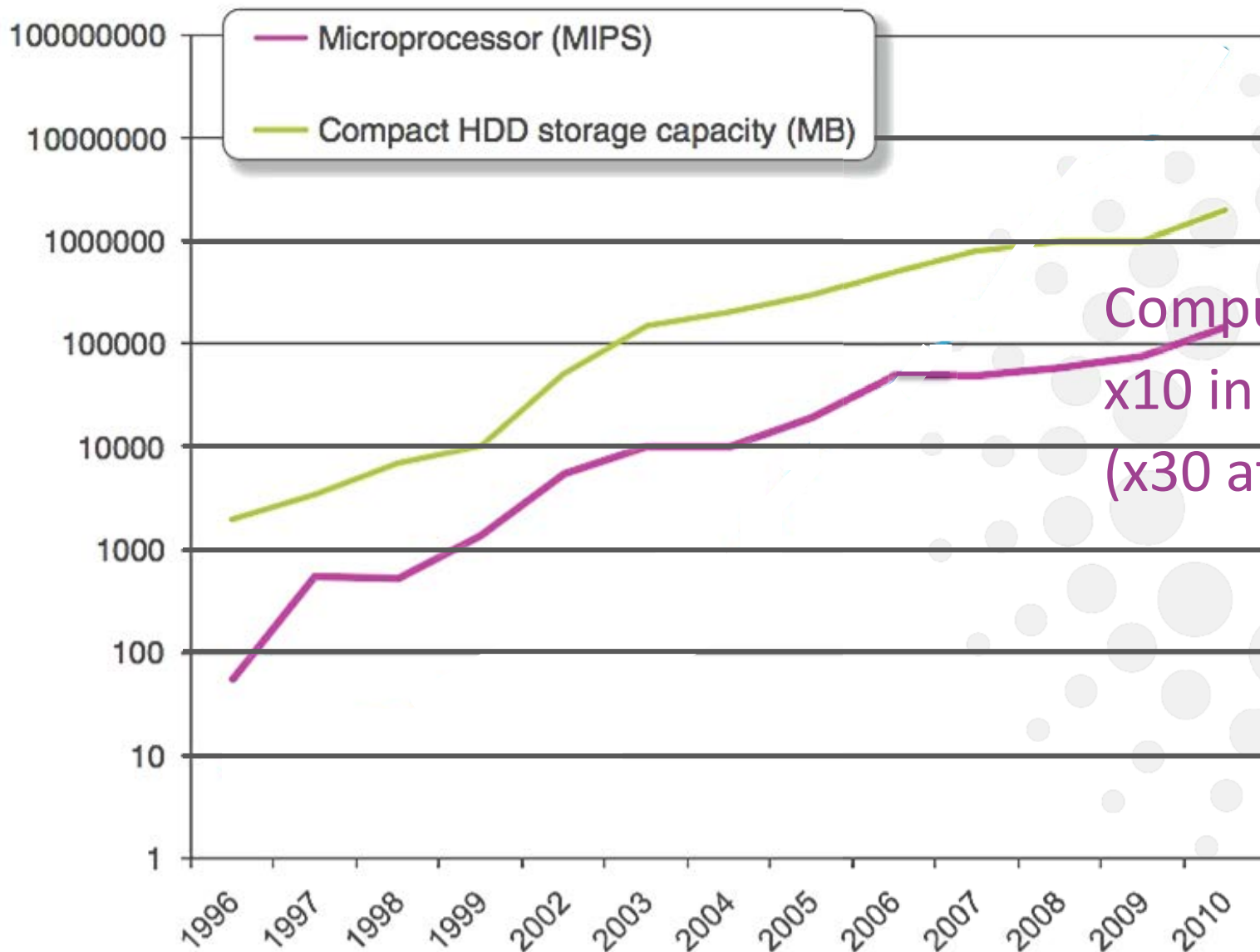


Transformative role of computation and 'big data'

Ian Foster
Director, Computation Institute

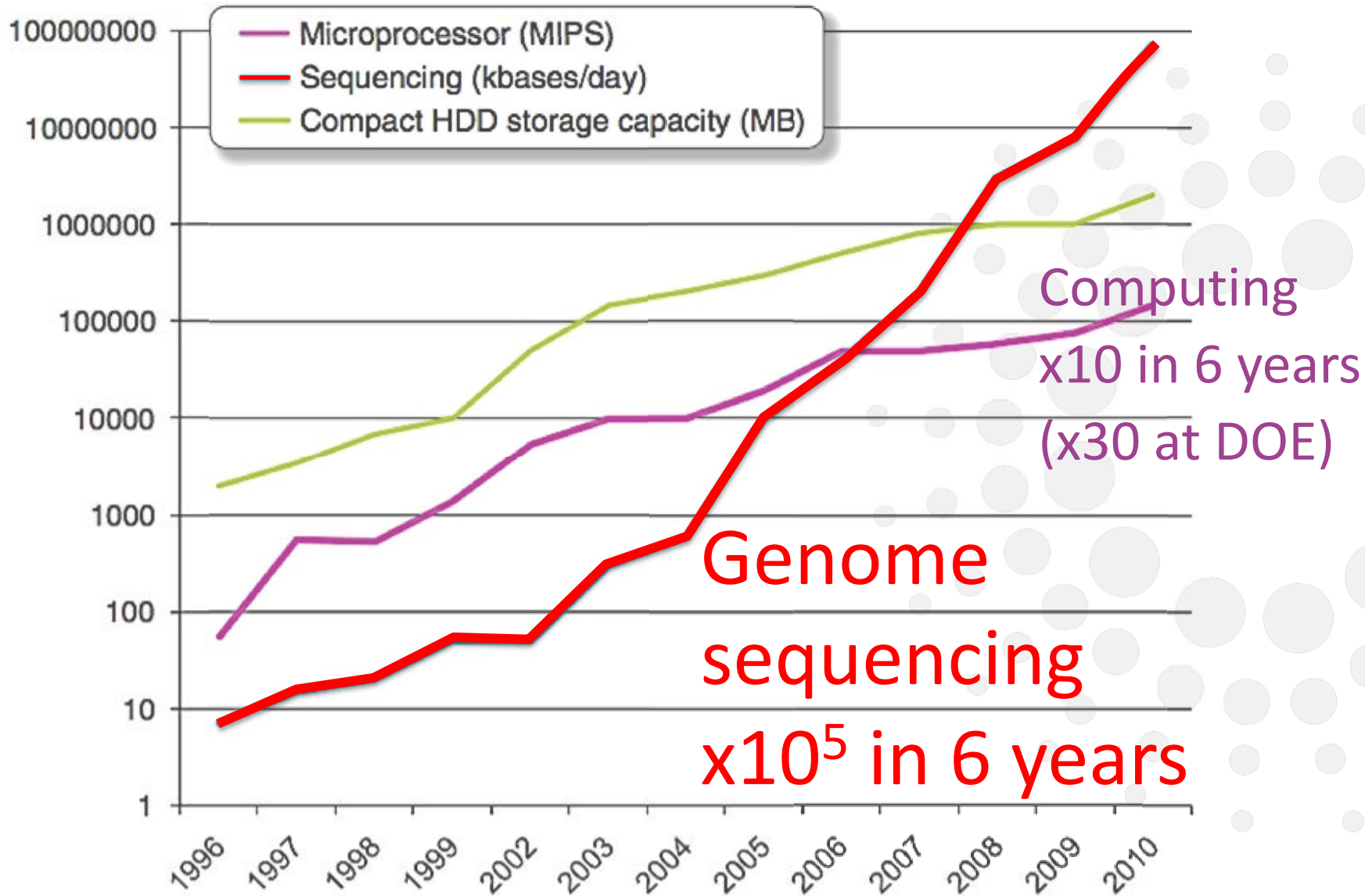
Secretary of Energy Advisory Board, April 17, 2012

Continued growth in computer power ...

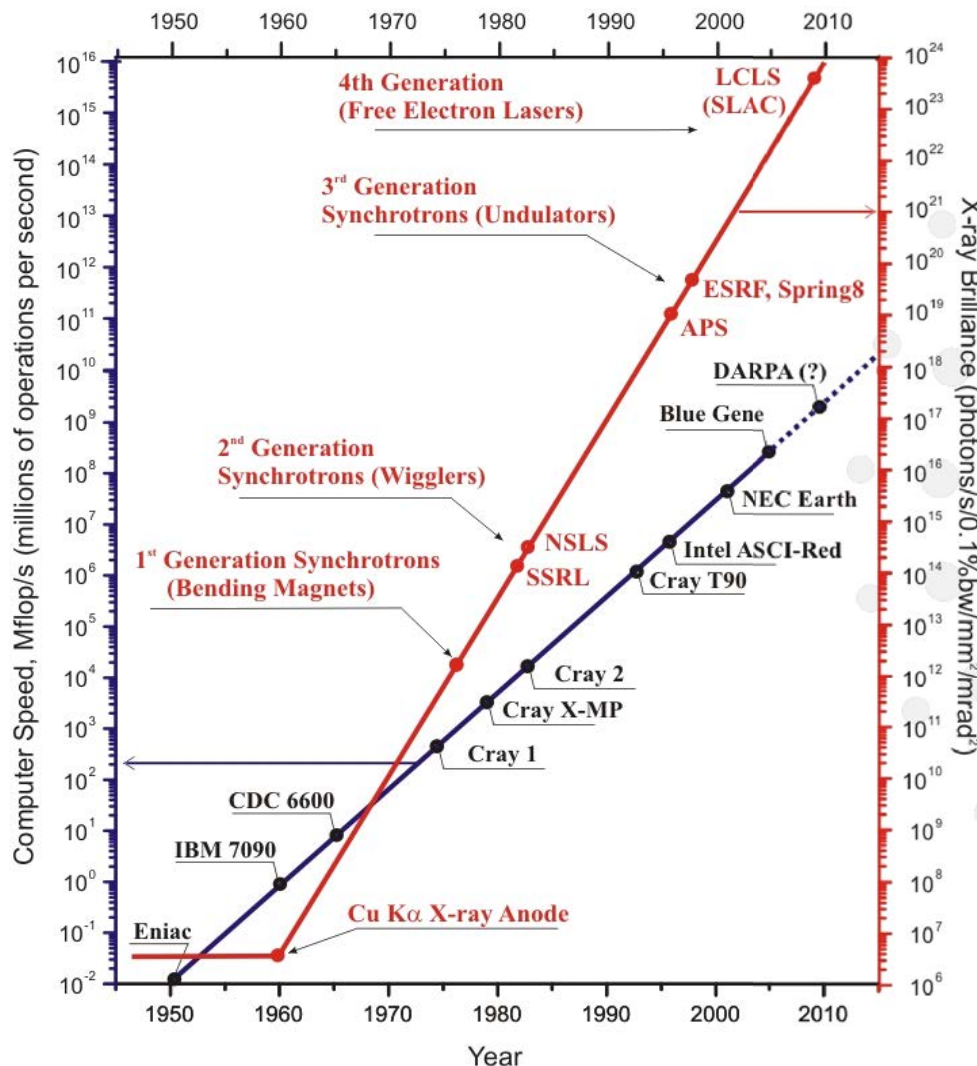


Computing
x10 in 6 years
(x30 at DOE)

... is dwarfed by increases in data



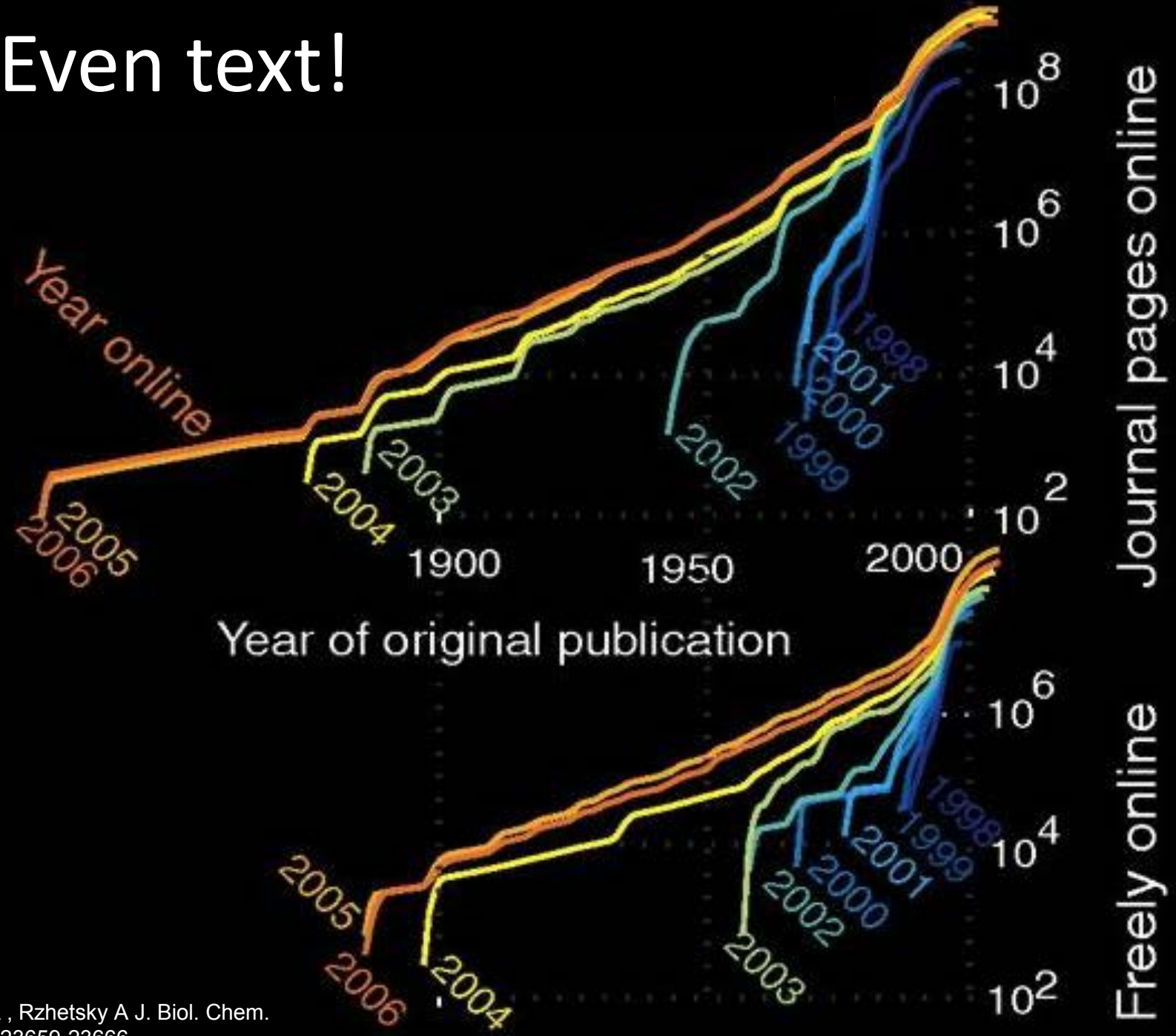
& in many other fields: e.g., X-ray sources



18 orders of magnitude in 5 decades!

12 orders of magnitude in 6 decades

Even text!





Imagine if, when faced with a problem, we could easily, both alone and within distributed teams:

- Assemble, integrate, and interpret all relevant data within a knowledge network
- Be informed of anomalies, patterns, and gaps
- Formulate and evaluate computational models
- Launch automated processes to test hypotheses, expand the knowledge network

All within an environment in which productive strategies can be easily scaled—and repeated

Discovery and innovation \propto

(People

x Simulation

x Data

x Process) Computation





Interdisciplinary
Science Teams
Understanding systems of systems



Home of the
Research Cloud
Amplifying human capabilities

Communication
Nexus
Exchange, education, engagement

The Computation Institute in context



Computation Institute

Biological
Sciences

Physical
Sciences

Computing,
Environment,
Life Sciences

Physical
Sciences and
Engineering

Humanities

Social
Sciences

Energy Eng.
and Systems
Analysis

Photon
Sciences

Law
school

Biz school

Harris
school

NORC

Oriental
Institute

Chapin
Hall

Fellows and projects



5

Computation Institute

Biological Sciences

13

Physical Sciences

24

Computing, Environment, Life Sciences

35

Physical Sciences and Engineering

9

Humanities

6

Social Sciences

5

Energy Eng. and Systems Analysis

8

Photon Sciences

Law school

Biz school

Harris school

NORC

Oriental Institute

Chapin Hall

The image shows two tall, dark grey server racks from IBM's Blue Gene Q series. The racks are filled with circuit boards and components, with some green indicator lights visible. The racks are positioned in a server room, with a tiled floor and other equipment visible in the background. The text is overlaid on the left side of the image.

Two Racks of IBM's Blue Gene Q
400 TFlops and 32,000 CPUs

MiRA at Argonne will be

48 Racks

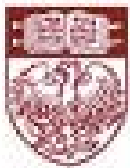

~800,000 CPUs

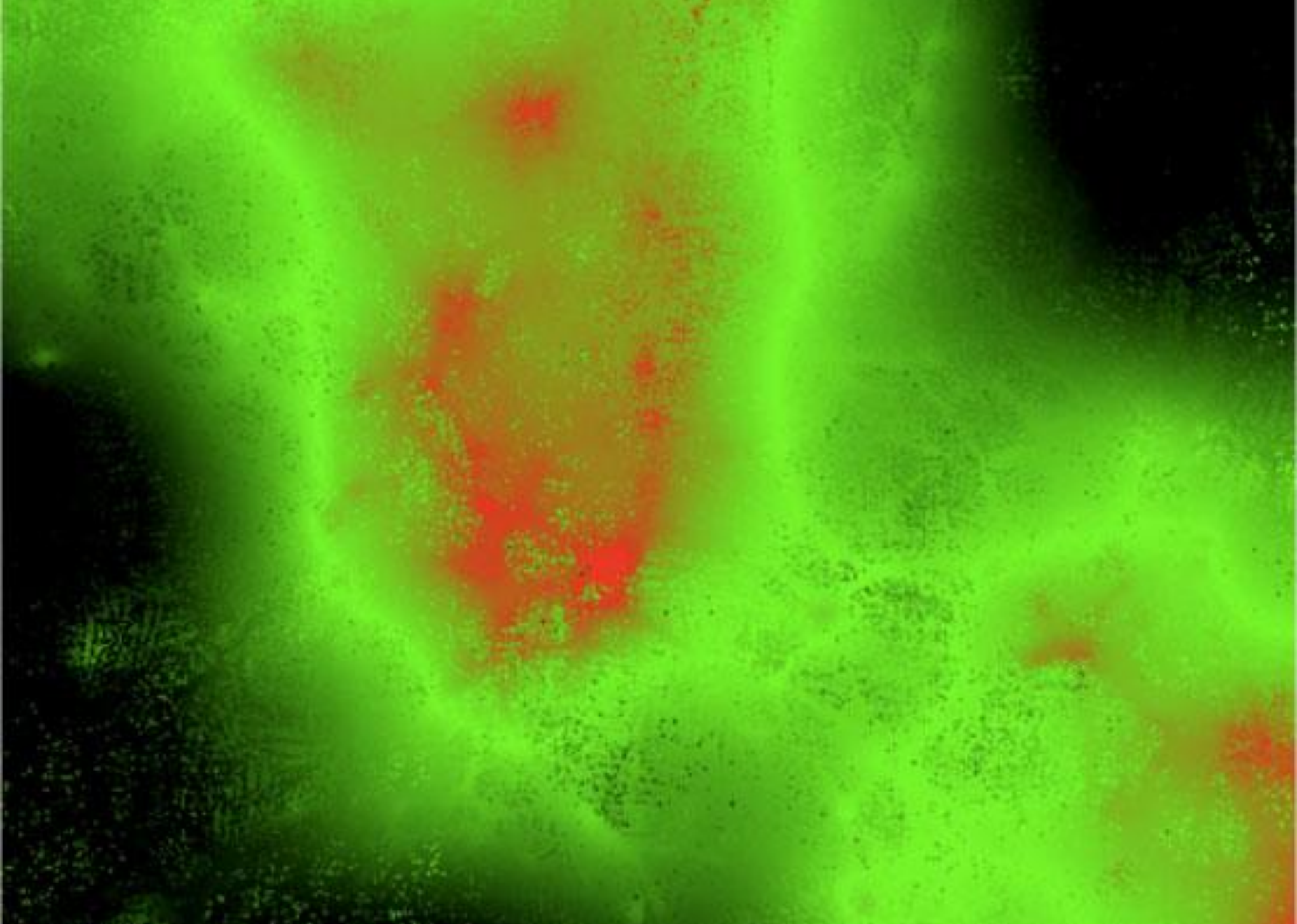
10 PFlops

35 Petabytes of storage

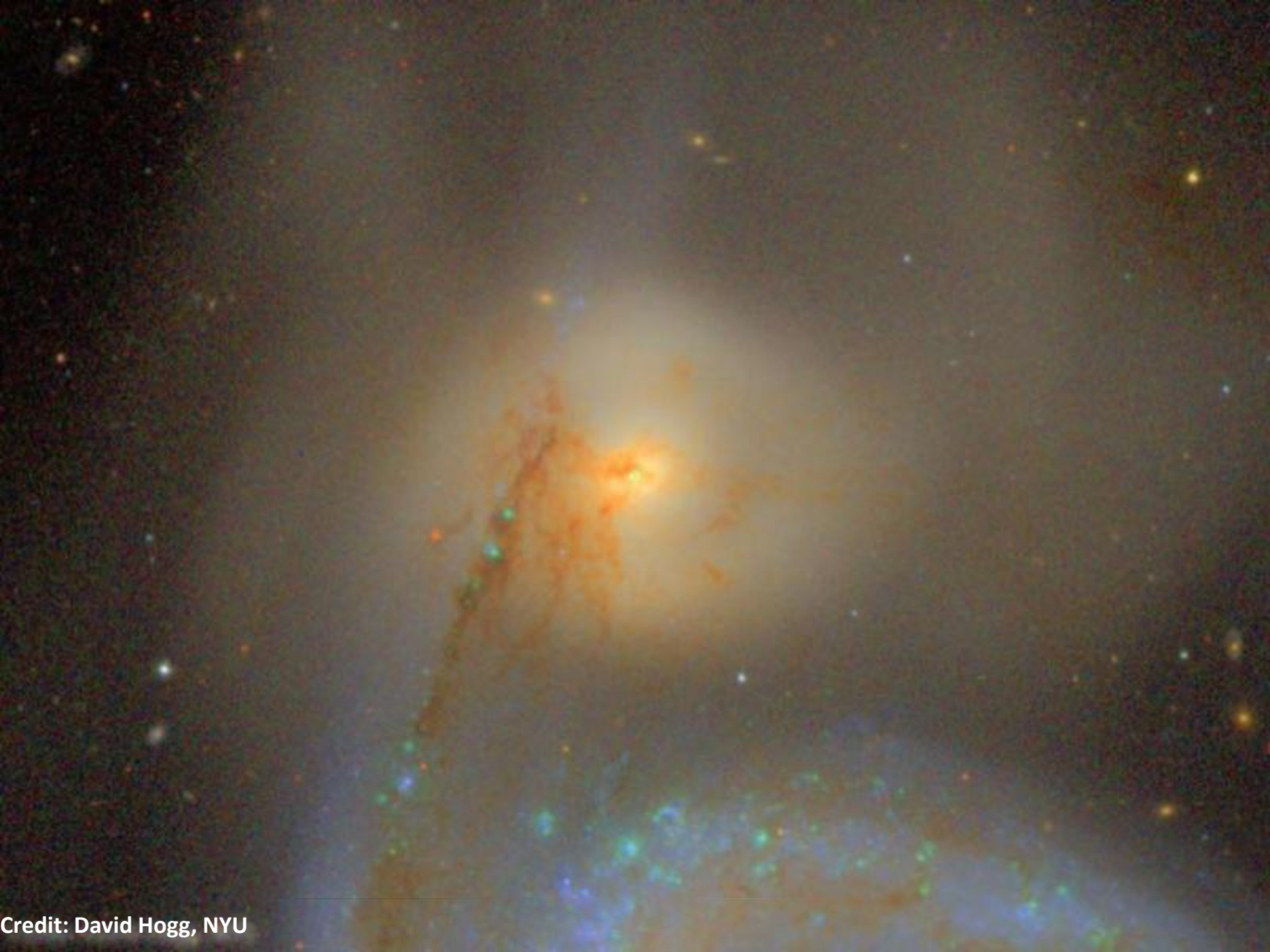
8 Megawatts of power



- **Alexei Khokhlov**, High-Speed Combustion and Detonation 
- **Gregory Voth**, Multiscale Molecular Simulations
- **Benoit Roux**, Large-Scale Simulations of Biomolecular Systems
- **Steve Pieper**, Ab-initio Reaction Calculations for Carbon-12 
- **Larry Curtiss**, Materials Design and Discovery: Catalysis, Energy Storage
- **Salman Habib**, Cosmic Structure Probes of the Dark Universe

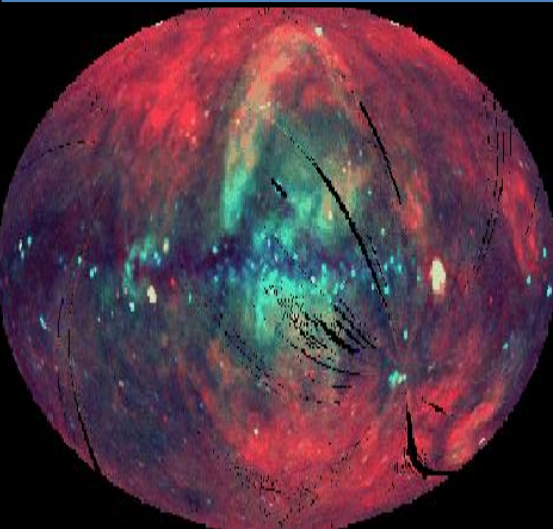


Matter distribution in the universe: S. Habib et al.

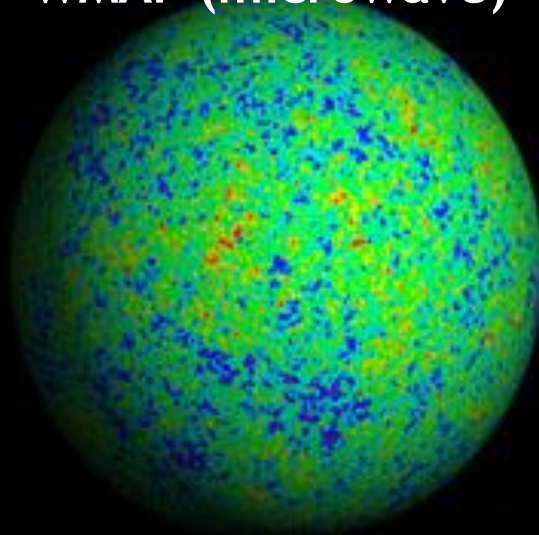


Credit: David Hogg, NYU

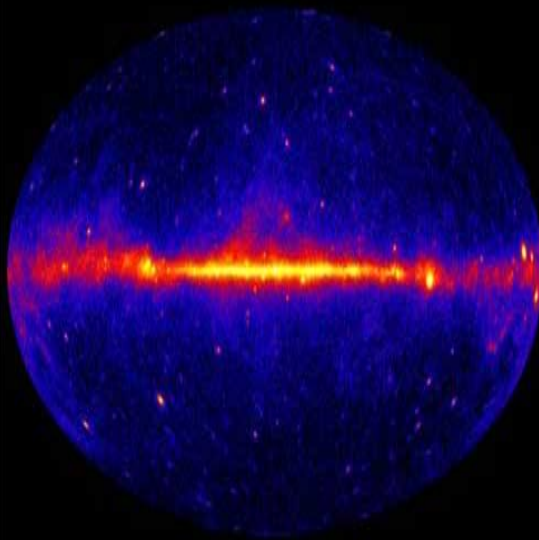
ROSAT (X-ray)



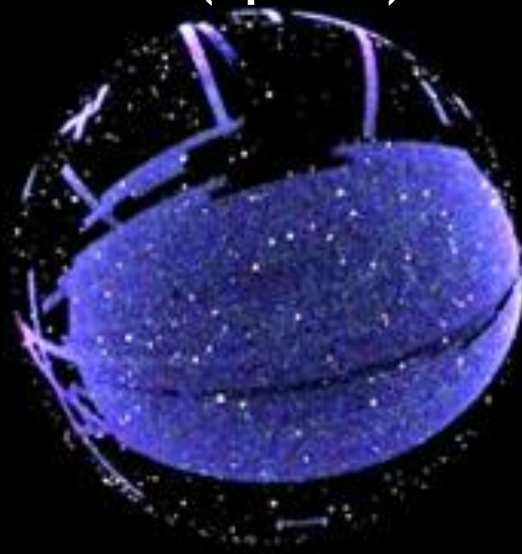
WMAP (microwave)



Fermi (gamma ray)



SDSS (optical)



MACHO et al.: 1 TB

Palomar: 3 TB

2MASS: 10 TB

GALEX: 30 TB

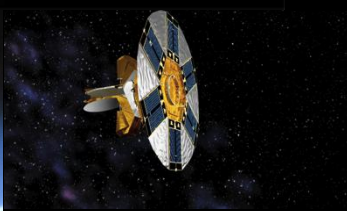
Sloan: 40 TB

Pan-STARRS:

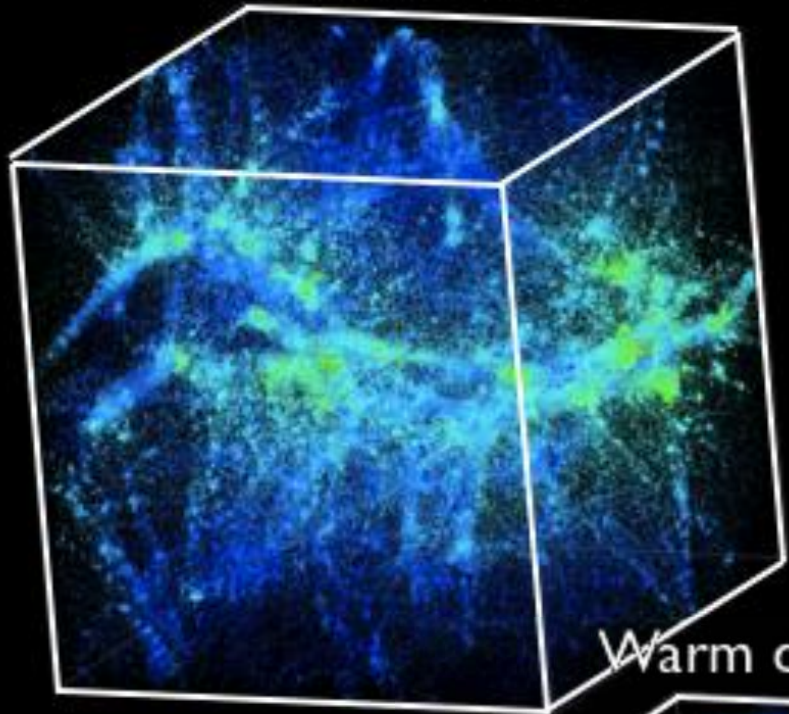
40,000 TB

LSST:

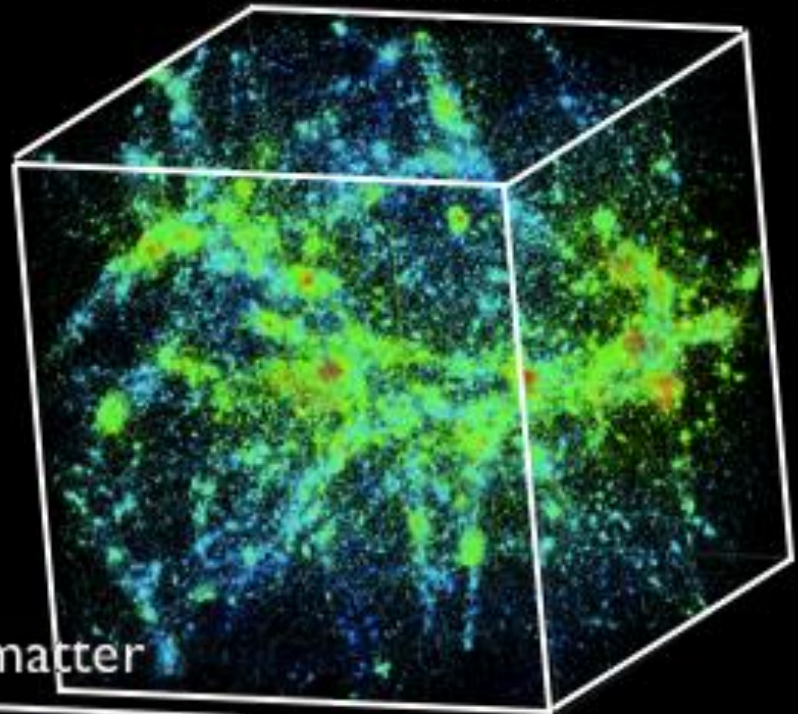
100,000 TB



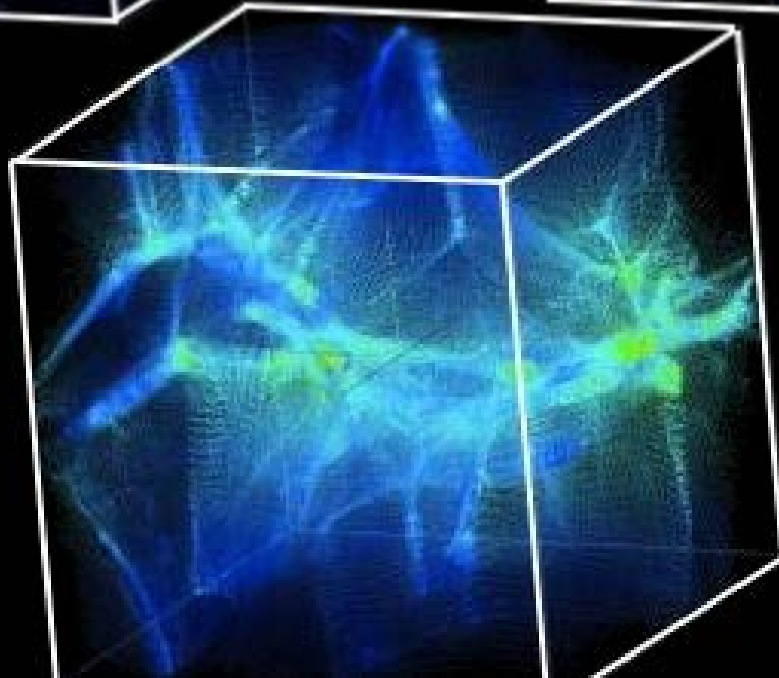
Standard Model

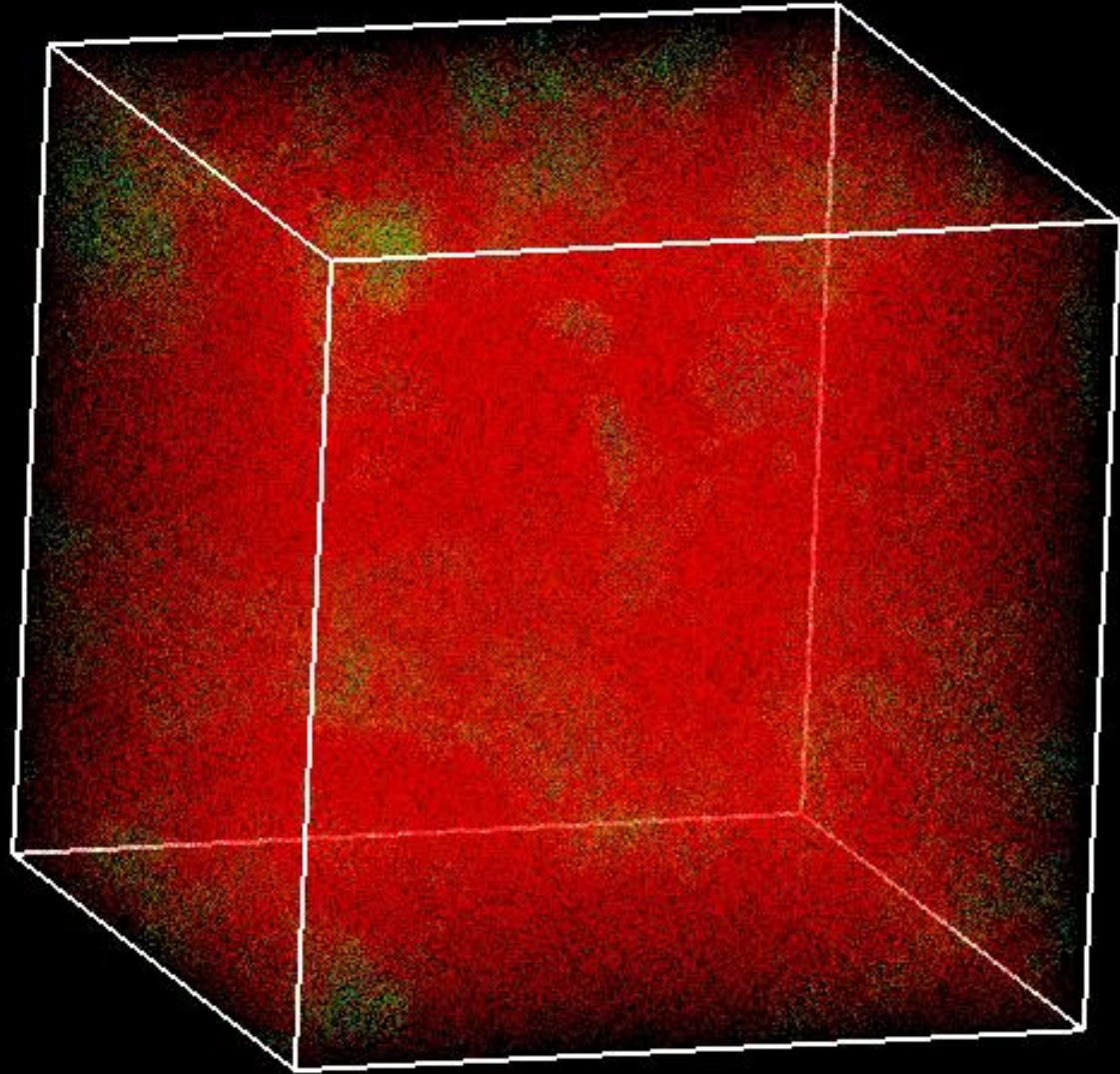


No dark energy



Warm dark matter





VELOCITY

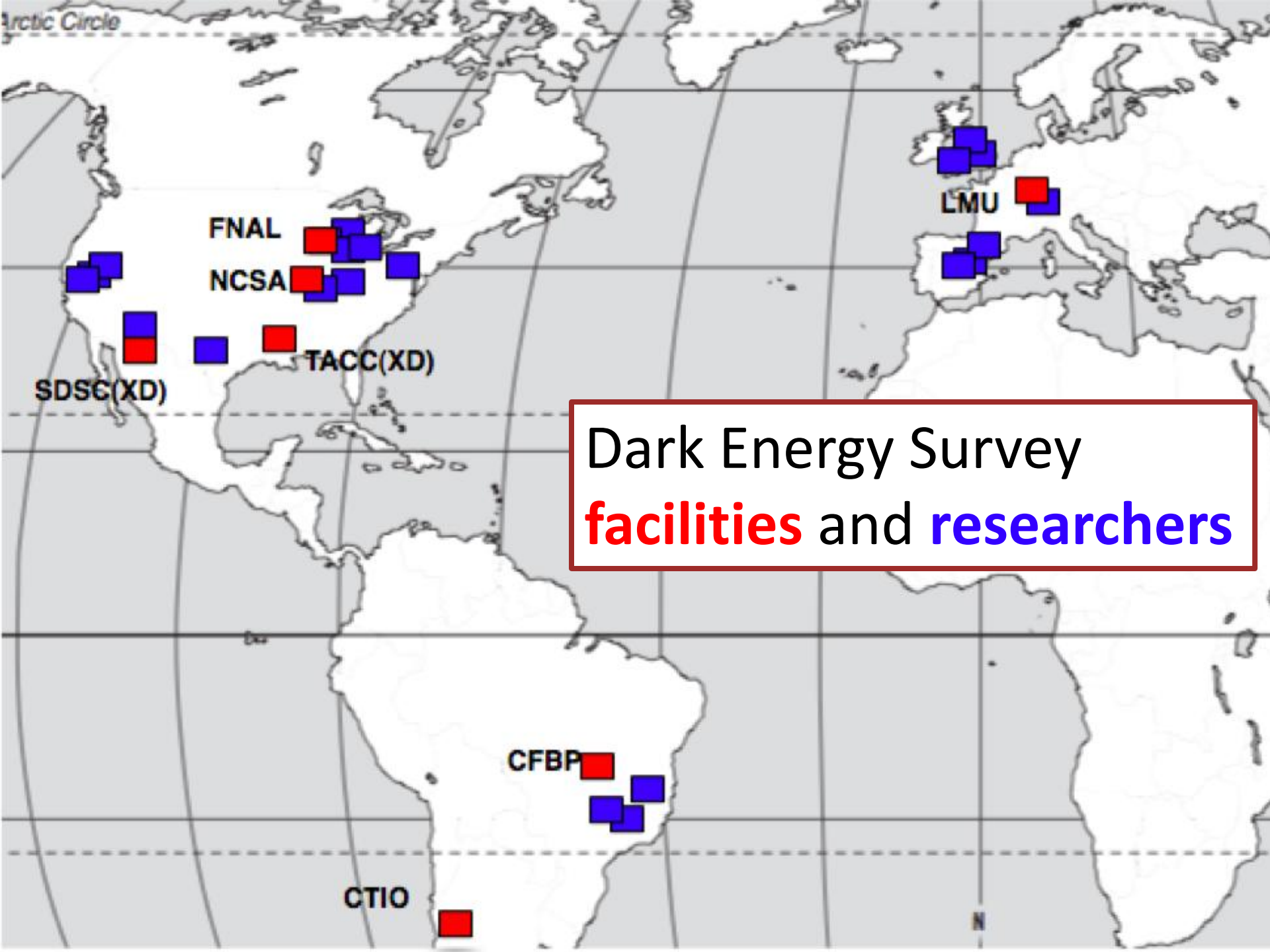
10000

7500

5000

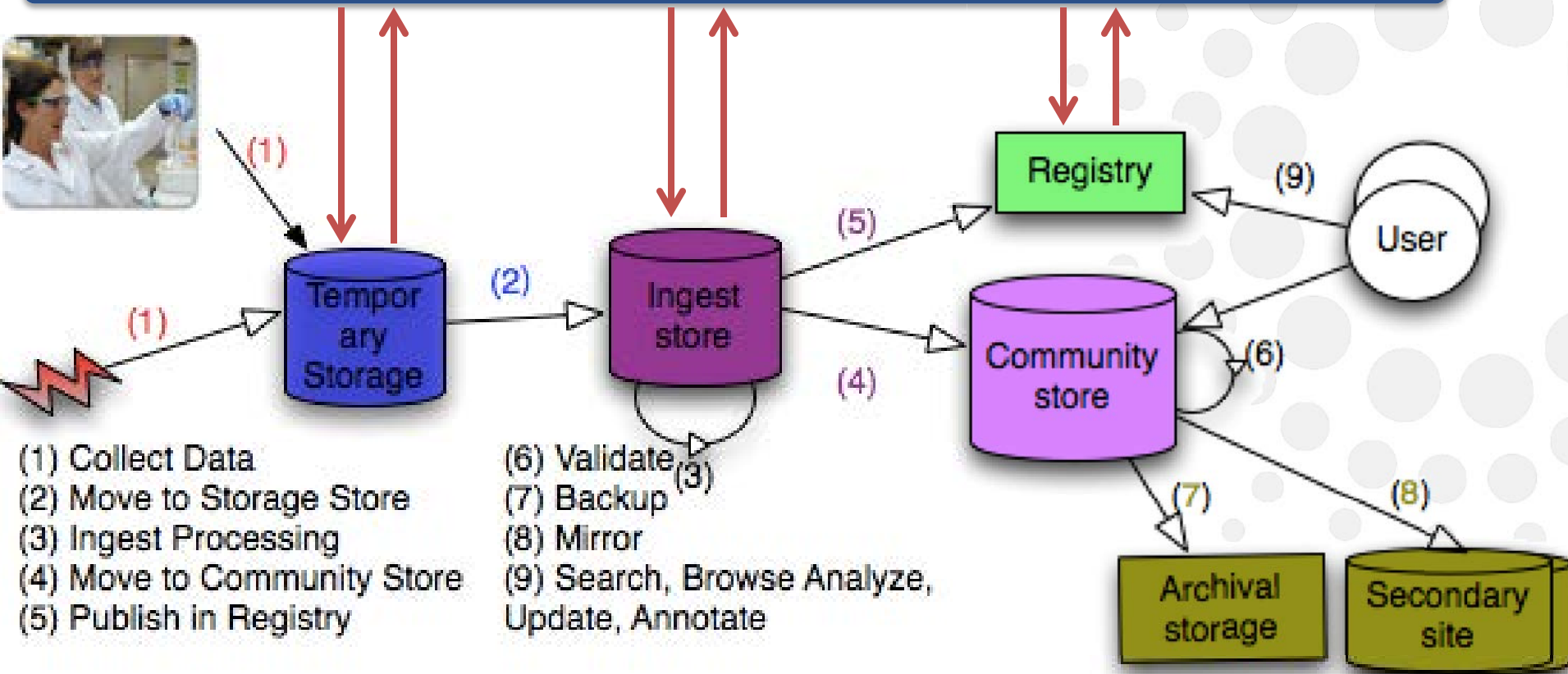
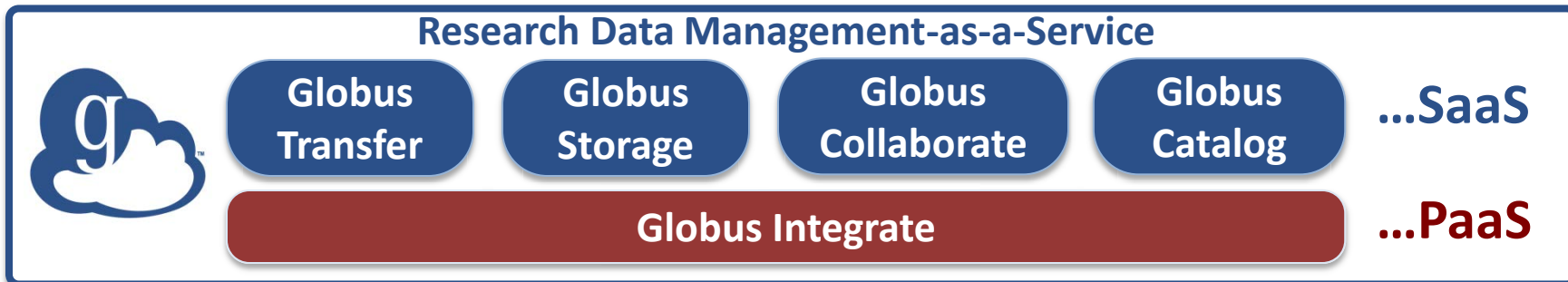
2500

0



Dark Energy Survey
facilities and **researchers**

Pervasive need for process automation



Moving 20 Terabytes LANL → Argonne



globus online Manage Data Groups News & Events About Support Log In Sign Up

Reliable, high-performance, secure file transfer.

Move files fast. No IT required.

+ WATCH A VIDEO
Globus Online in a nutshell

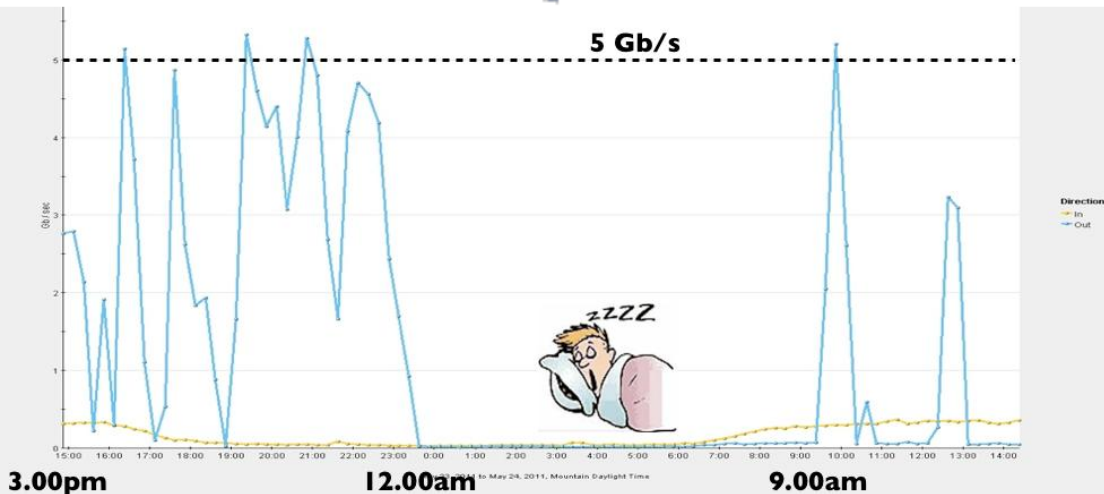
→ GET STARTED
Sign up and get moving

3,662,210,979 MB TRANSFERRED

Why Use Globus Online? See how easy file transfer can be

For HPC Resource Owners Enable Globus Online for your users

For Developers Integrate with Globus Online



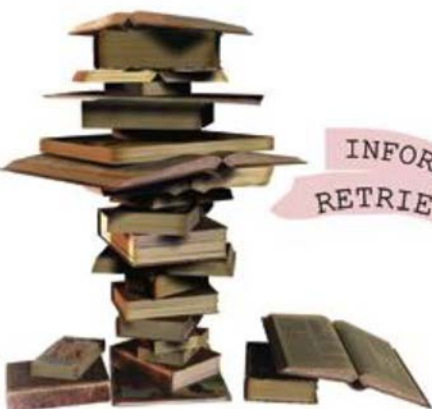
Exploring science via text mining



Extracting relations
between entities in a large corpus,
mapping what is known, and
generating new hypotheses



Andrey Rzhetsky



INFORMATION
RETRIEVAL

NAMED ENTITY
RECOGNITION



Gtf is an abbreviation for glycosyltransferase
O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.
The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer.
During malignant transformation, glyco-epitopes of MUC1 become exposed.
O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.
O-GalNAc modified peptides are resistant to proteolysis
Nishitani et al. 2001. OGT inhibitor

INFORMATION
EXTRACTION



What can text mining do for us?



Match problems with solutions discovered in other fields

Generate unexpected hypotheses

Enable probabilistic reasoning: finding consistent and conflicting cliques

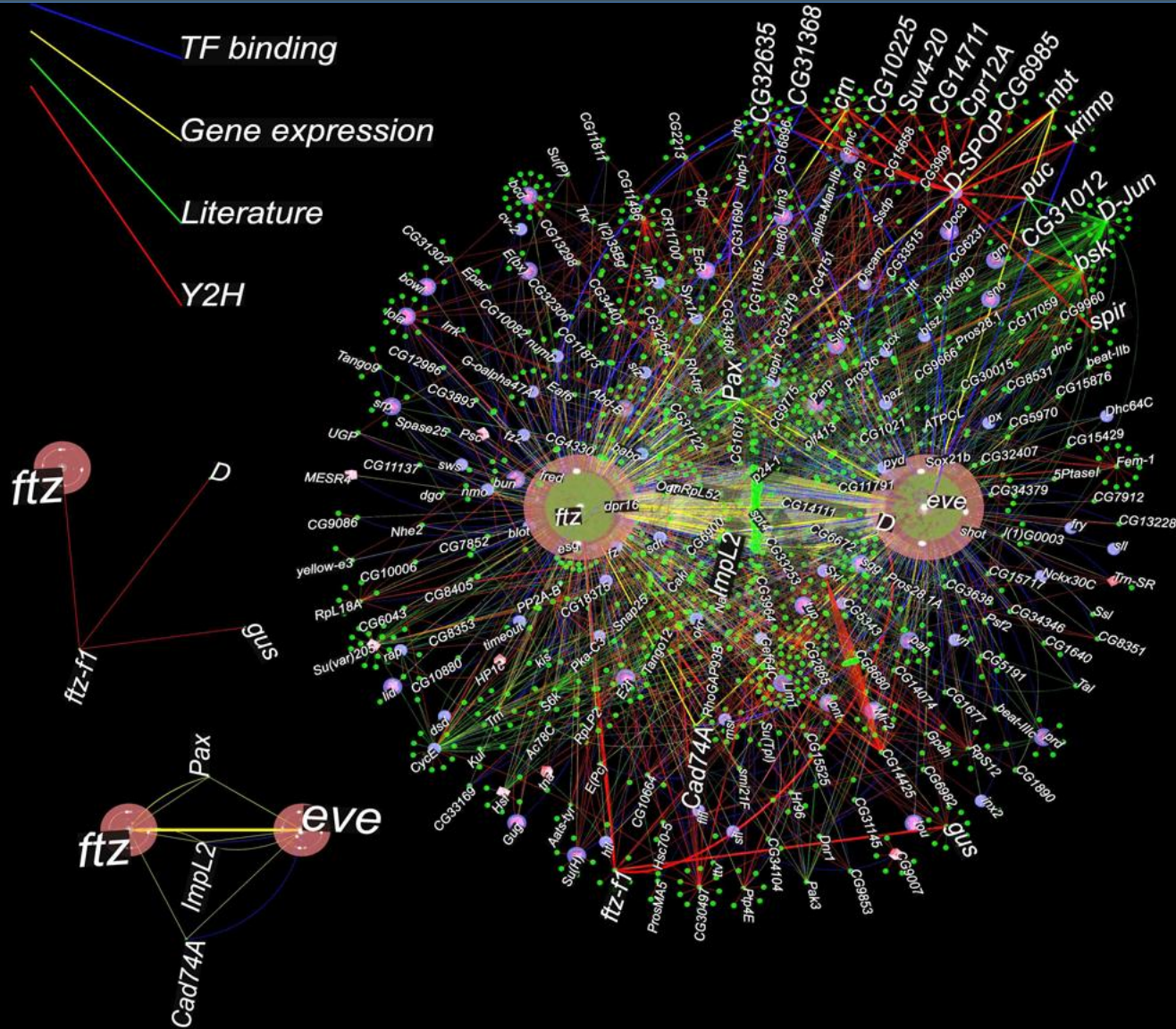
Create a real-time map of a scientific field

Challenge dynamics of scientific beliefs

Fly networks → human cancer phenotypes

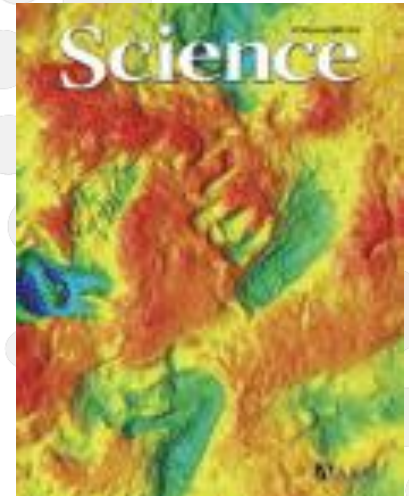


Analysis of:
>3,000 genes
4 data types
~6,000 edges



Analysis of *Drosophila* Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer

Jiang Liu, Murad Ghanim, Lei Xue, Christopher D. Brown, Ivan Iossifov, Cesar Angeletti, Sujun Hua, Nicolas Nègre, Michael Ludwig, Thomas Stricker, Hikmat A. Al-Ahmadie, Maria Tretiakova, Robert L. Camp, Montse Perera-Alberto, David L. Rimm, Tian Xu, Andrey Rzhetsky, and Kevin P. White



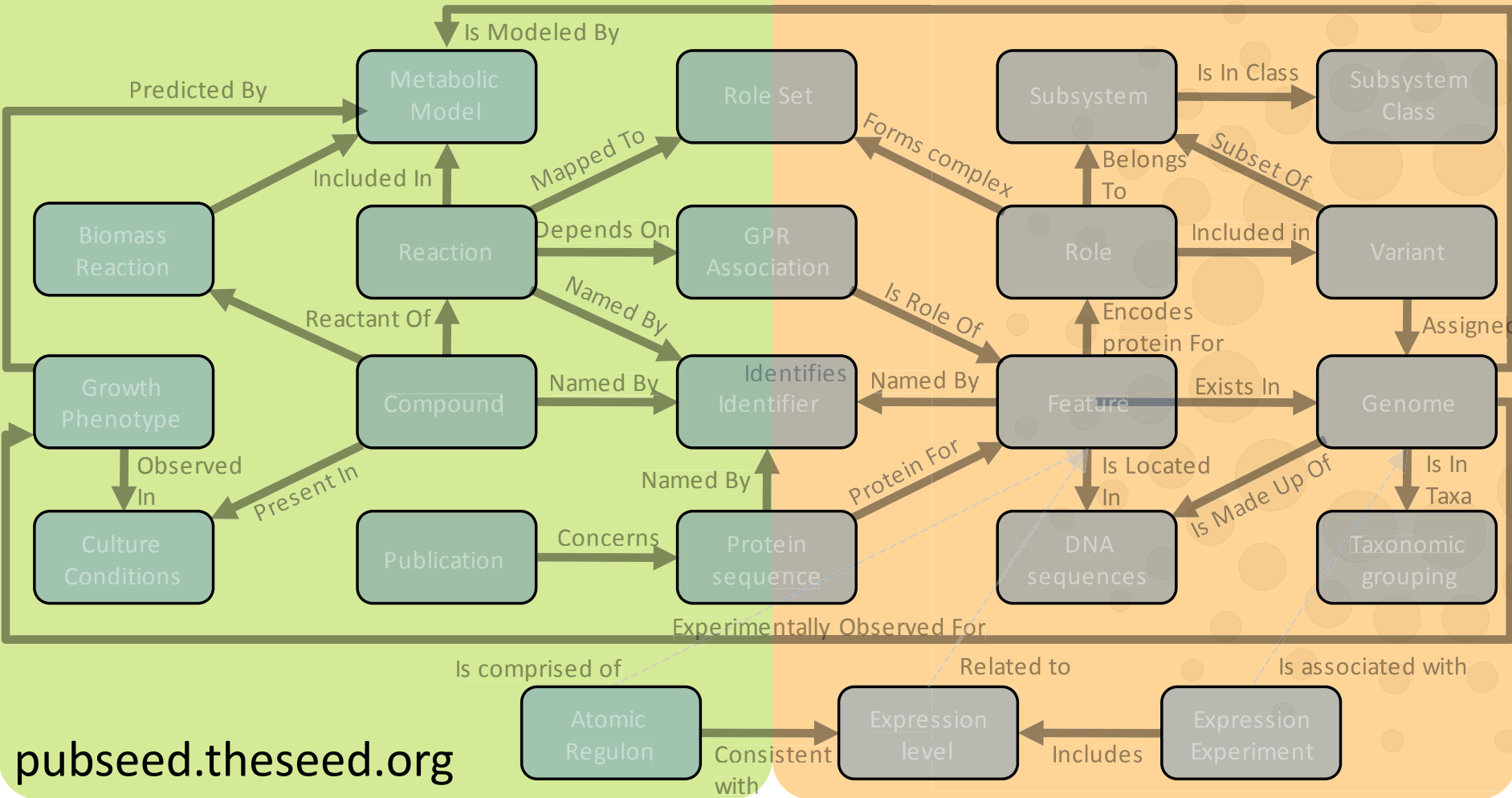
Science 27 February 2009: 1218-1222.
DOI: 10.1126/science.1157669

SEED/Model SEED database



Model SEED

SEED



pubseed.theseed.org



SEED/RAST

- 3,000+ publically available genome annotations
- **30,000+ private genomes annotated**
- 1000 subsystems
- 14,000 functional roles

Model SEED

- 3000+ publically available metabolic models
- **10,000+ private models constructed**
- 13,000 reaction
- 16,000 compounds
- 589 media conditions

MG-RAST

metagenomics analysis server



LOGIN

REGISTER

PASSWORD

FORGOT?

login



Browse Metagenomes

search for metagenomes



Register



Contact



Help



Upload*



News

About

MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

# of metagenomes	44,342
# base pairs	12.1 Tbp
# of sequences	111.98 billion
# of public metagenomes	7,694

The server provides web based upload, quality control, automated annotation and analysis for samples up to 10GBp. Comparison between large numbers of samples is enabled via pre-computed abundance profiles.

* login required



EPA Preliminary Analysis of the Waxman-Markey Discussion Draft

Household consumption is reduced by 0.02-0.11% in 2015 and 0.17-0.19% in 2020 and 0.37-0.39% in 2030, relative to the no policy case.

Household consumption under the WM Draft scenario still increases by 9-10% percent between 2010 and 2015 and 18-19% between 2010 and 2020.



ADAGE (RTI Inc.)

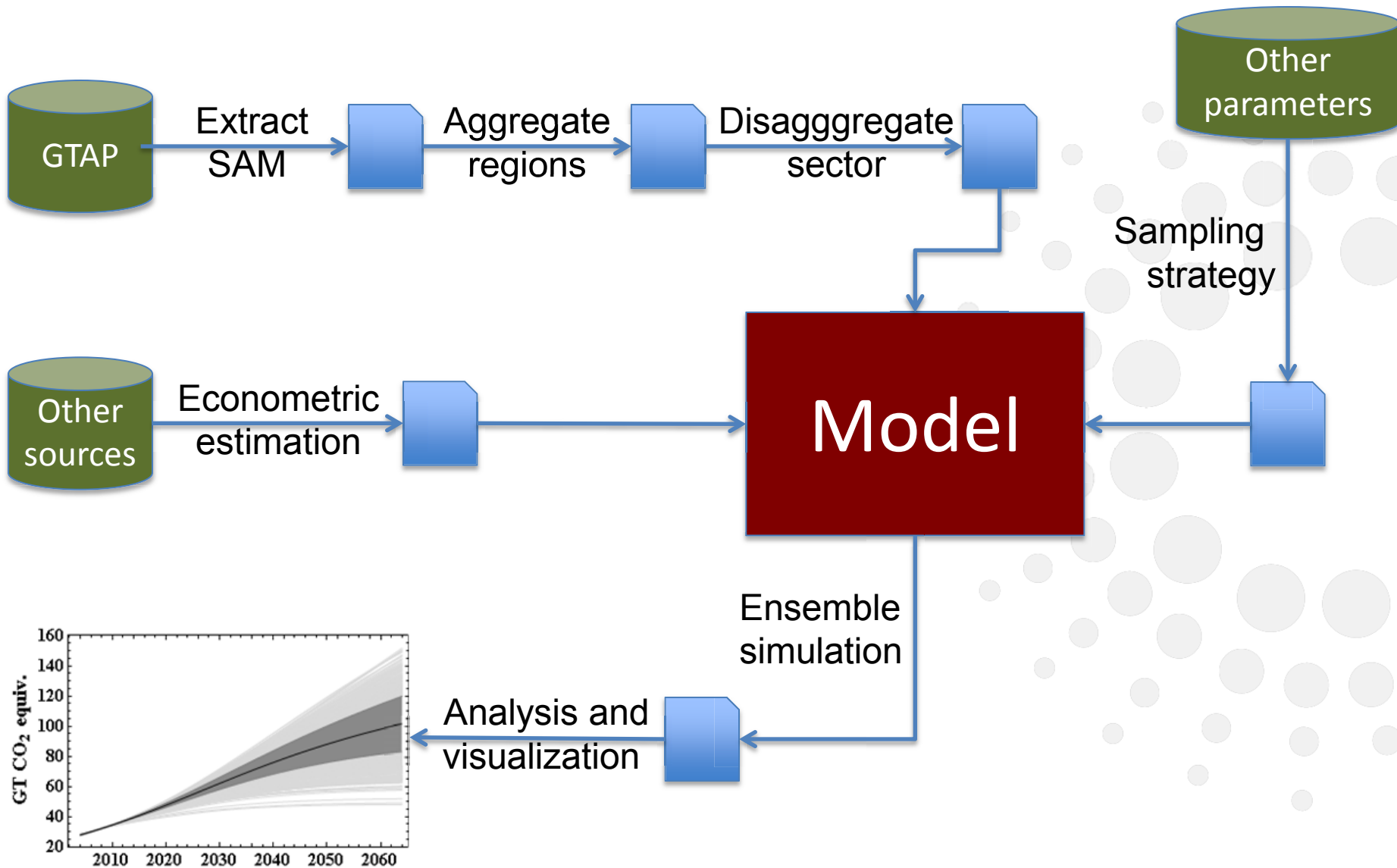
IGEM (Jorgenson Assoc.)

IPM (ICF Consulting)

FASOM (Texas A&M)

Four closed models

Open Source CIM-EARTH Framework



A future energy research environment



Data
in



Programs
& rules in



Results
out



“No limits”

- Storage
- Computing
- Format
- Program

Weighted							
HI	min	max	mean	std dev	count	median	
0	0	108087	22708	15364	923	19943	
1	0	34808	19509	9442	10	21786	
2	0	25795	17005	8115	7	18656	
3	14464	30102	21355	7982	3	19500	
4	9427	59183	33118	13457	21	30795	
5	4999	42458	21992	8208	39	21176	
6	10857	22386	17711	4565	9	20400	
7	9418	30347	21344	8679	9	26197	
Non-weighted							
HI	min	max	mean	std dev	count	median	
0	0	108087	22979	15364	863	20208	
1	0	64962	21247	14022	76	19932.5	
2	8770	55315	22046	11355	22	19054	
3	0	33886	21825	10862	14	23600.5	
4	15383	40588	28753	12672	3	30288	
5	8414	42458	21541	8094	29	20833	
6	4999	37051	21706	8457	12	20502	
7	7927	26821	17374	13360	2	17374	

Allowing for

- Versioning
- Provenance
- Collaboration
- Annotation



Thank you!

foster@anl.gov