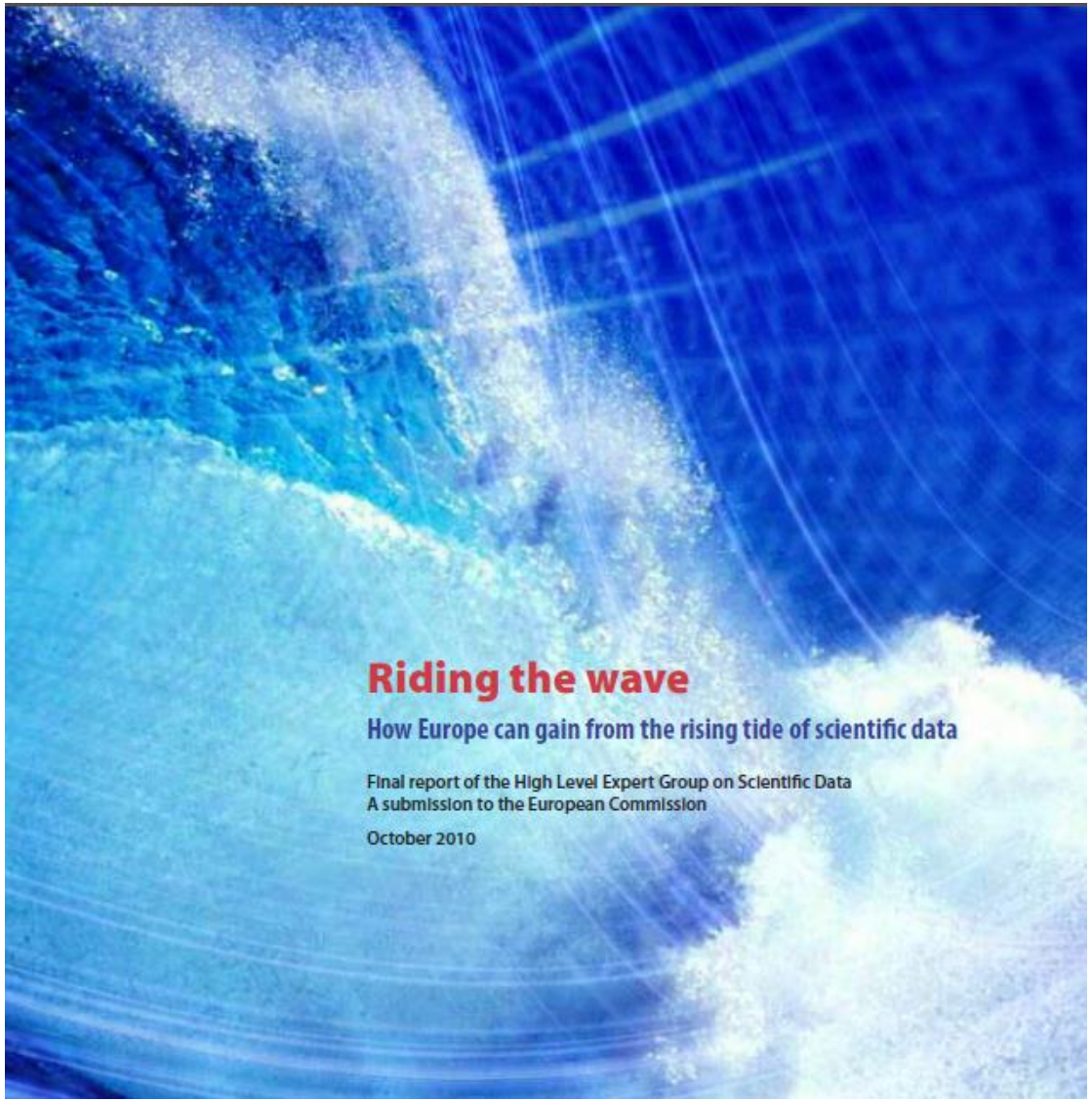
The Fourth Paradigm: Data-Intensive Scientific Discovery

Tony Hey

Chief Data Scientist

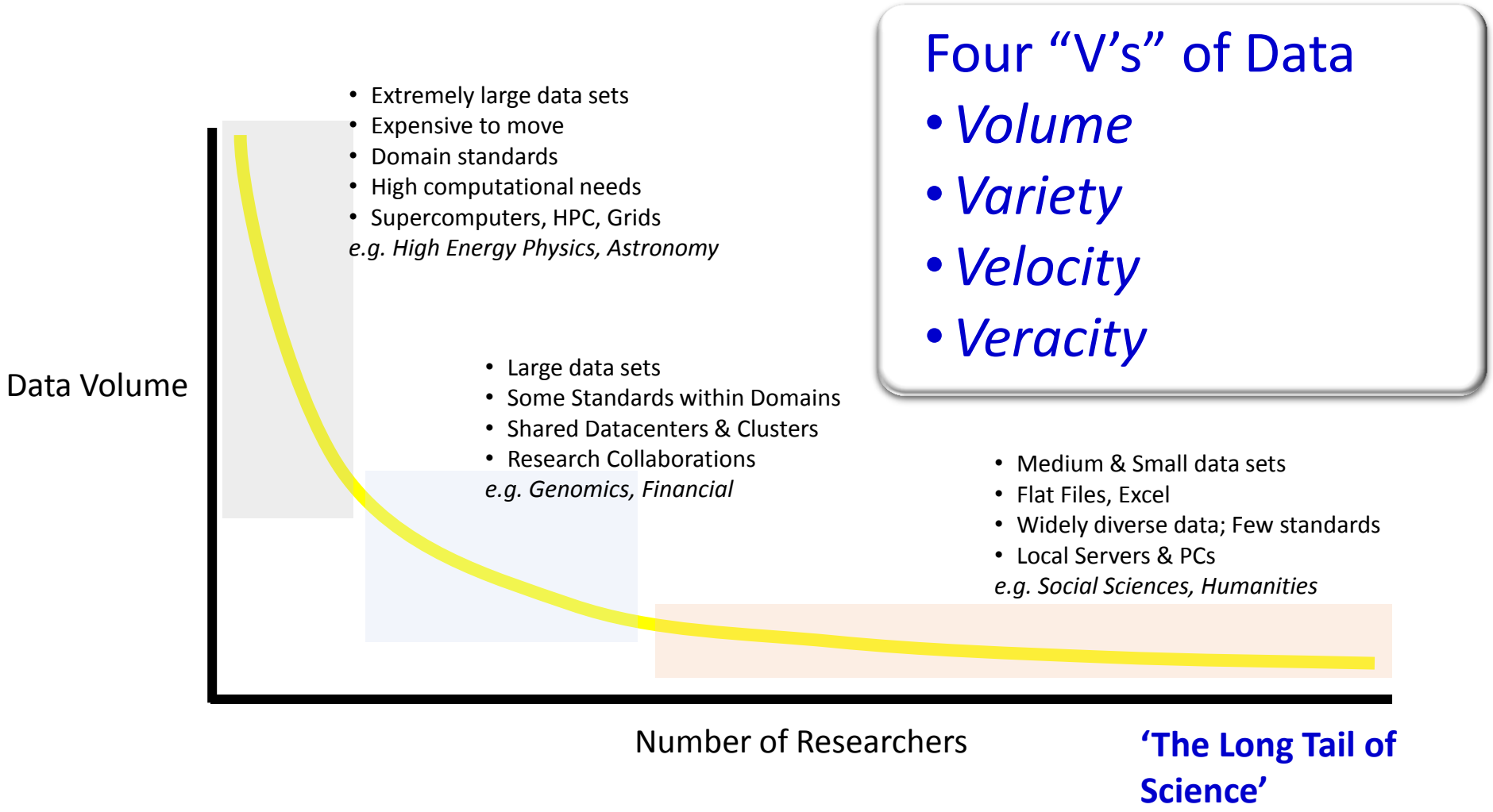
STFC

The Data Deluge – Data-Intensive Science



http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Much of Science is now Data-Intensive



The 'Cosmic Genome Project': The Sloan Digital Sky Survey

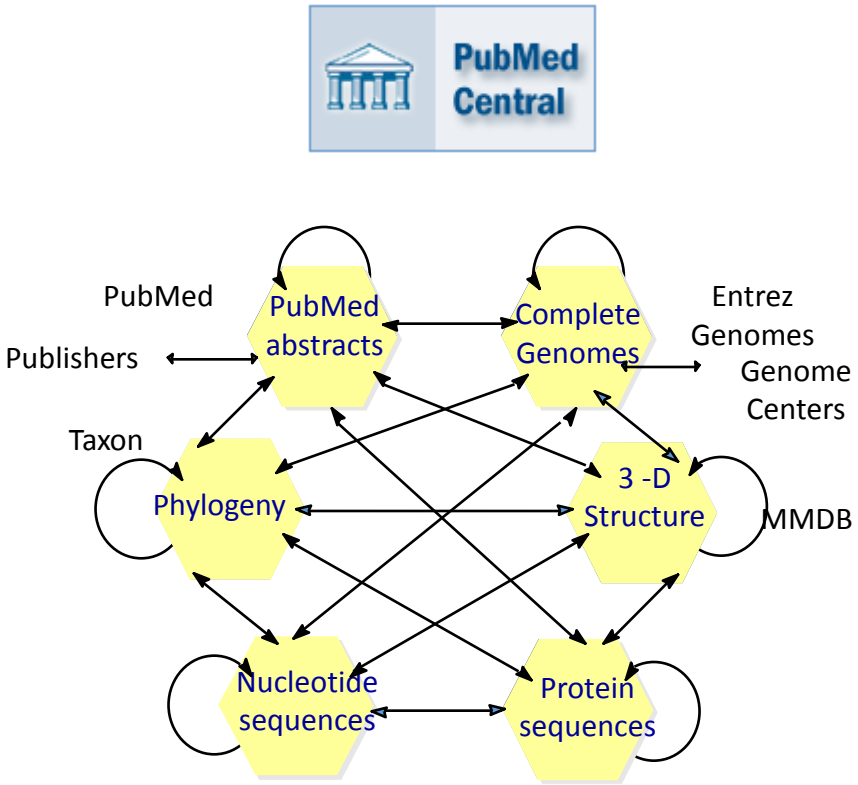
- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images
  - 40 TB of raw data => 120TB processed data
  - 5 TB catalogs => 35TB in the end
- Started in 1992, 'finished' in 2008
  - ➢ SkyServer Web Service built at JHU by team led by Alex Szalay and Jim Gray

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA

The US National Library of Medicine

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.



Entrez cross-database search

e-Science and the Fourth Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

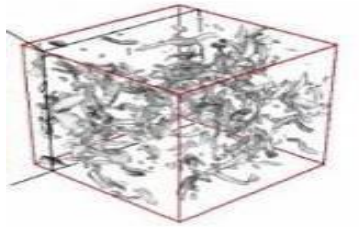
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

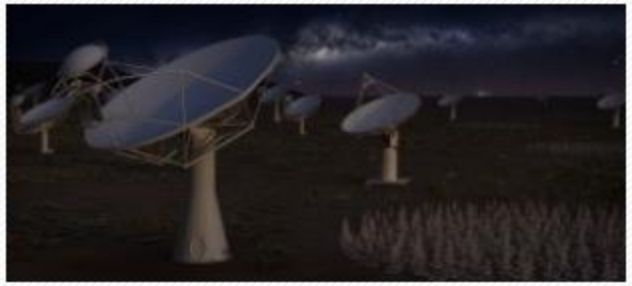


eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

With thanks to Jim Gray

SKA Project



Artist impression of the Square Kilometre Array

The Square Kilometre Array (SKA) project is an international effort to build the world's largest radio telescope, with eventually over a square kilometre (one million square metres) of collecting area. The scale of the SKA represents a huge leap forward in both engineering and research & development towards building and delivering a unique instrument, with the detailed design and preparation now well under way. As one of the largest scientific endeavours in history, the SKA will bring together a wealth of the world's finest scientists, engineers and policy makers to bring the project to fruition.

Latest News

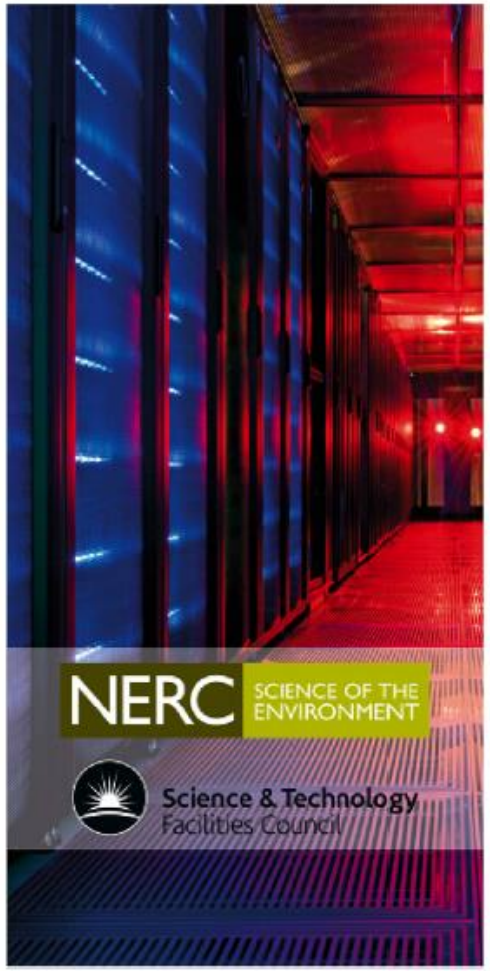
22nd December 2015
2015: a big year for ASKAP!

21st December 2015
Outcomes Of The 19th SKA Board Meeting

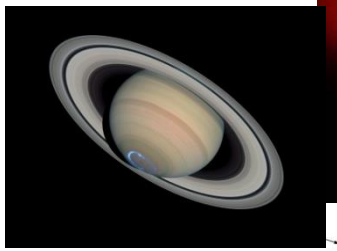
7th December 2015
Australia Announces AUS$293.7 Million for the SKA

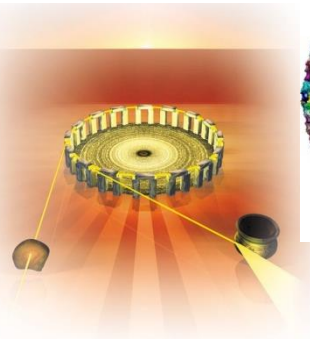
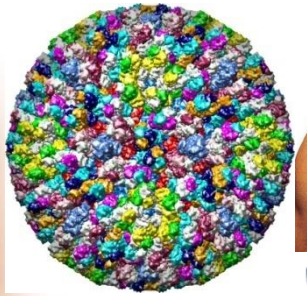
Centre for Environmental Data Analysis: JASMIN infrastructure

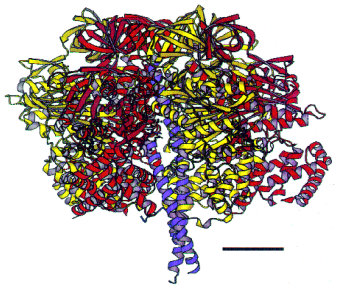
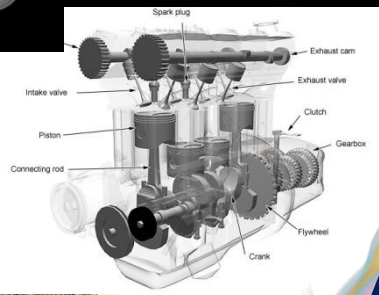
Part data store, part supercomputer, part private cloud…



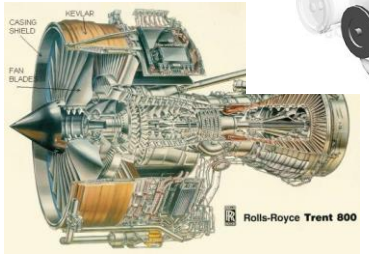
- 16 PB Fast Storage (Panasas, many Tbit/s bandwidth)
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.
- Some high memory nodes, a range, bottom heavy.



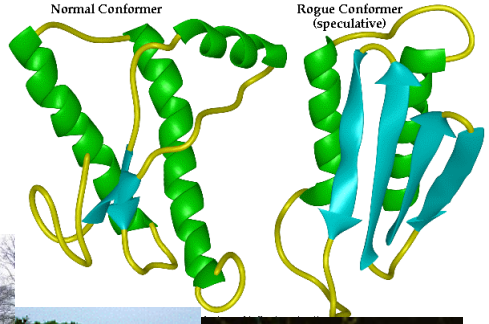



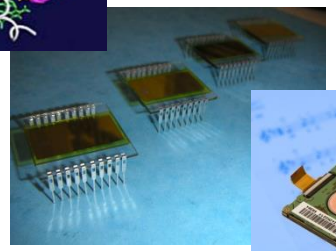




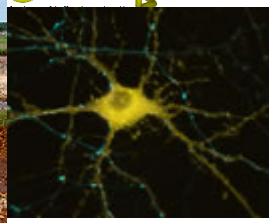
Applications of Synchrotron Radiation







e-Infrastructure and Experimental and Observational Data (EOD)

UK e-Science Program: Six Key Elements for a Global e-Infrastructure (2004)

1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centres for Open Software
4. Technologies and standards for Data Provenance, Curation and Preservation
5. Open access to Data and Publications via Interoperable Repositories
6. Discovery Services and Collaborative Tools

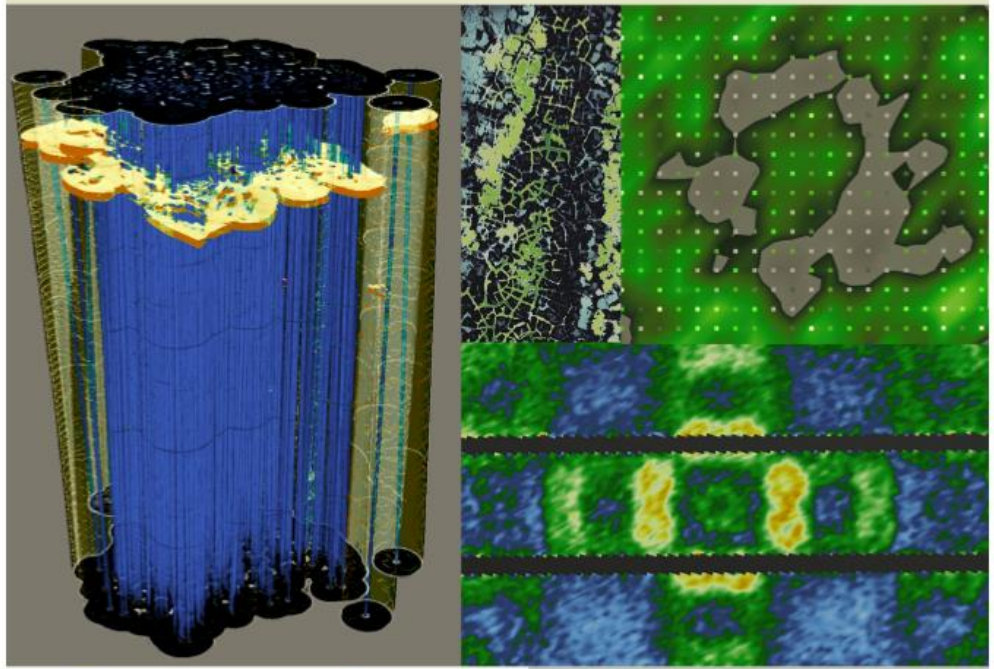
Plus:

Supercomputing and HPC resources

Training of Scientific Software Engineers and Data Scientists

Report of the
DOE Workshop on

Management, Analysis, and Visualization of Experimental and Observational Data
The Convergence of Data and Computing



U.S. DEPARTMENT OF ENERGY | Office of Science

September 29th - October 1, 2015
Bethesda, MD

Prepublication Copy—Subject To Further Editorial Correction

Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020

Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

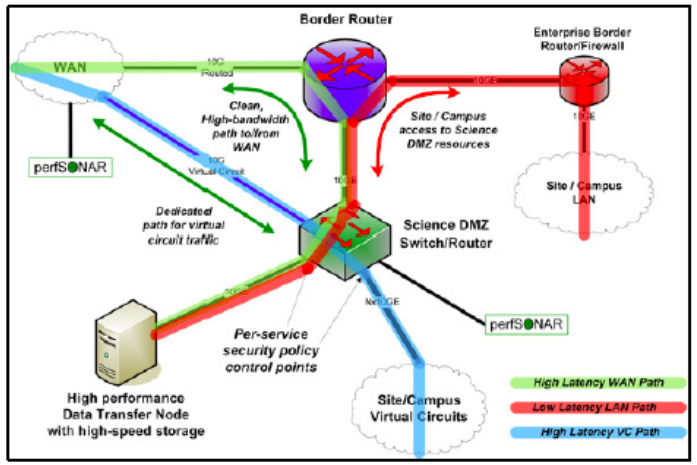
The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Science DMZ, a network design pattern, improves the baseline end-to-end performance through ongoing global adoption

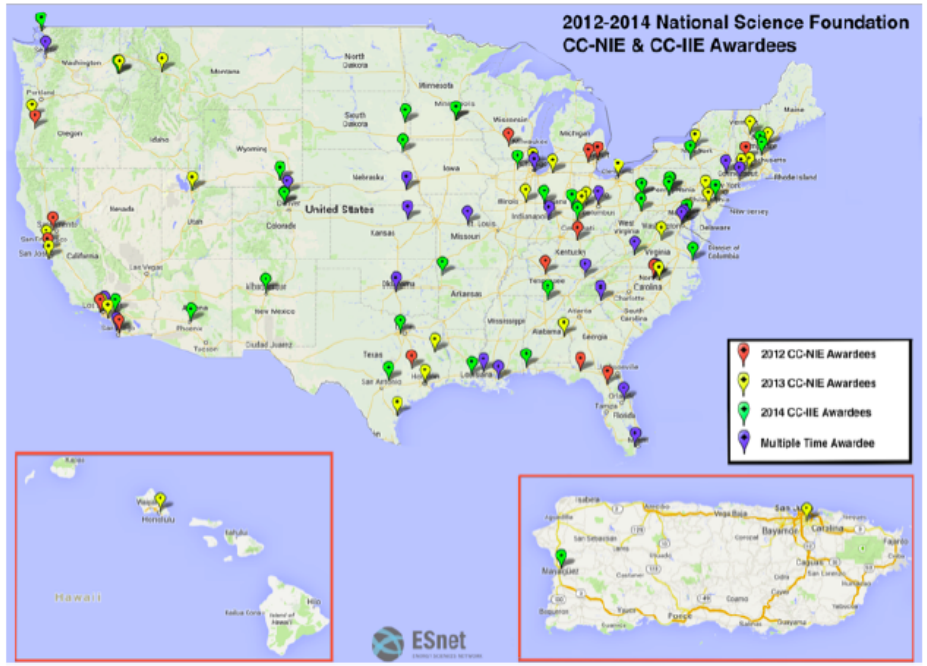
Science DMZ, facilitating great end-to-end network hygiene

- "Friction free" network path
- Dedicated, high-performance Data Transfer Nodes (DTNs)
- Performance measurement/test node

A **prerequisite** for any superfacility architecture



Science DMZ design pattern



$80M+ funding to implement Science DMZ design pattern in Universities

SPOT SUITE TRANSFORMS BEAMLINE SCIENCE

SPOT Suite brings advanced algorithms, high performance computing and data management to the masses

AUGUST 18, 2014 | Tags: Accelerator Science, Carver, Data Transfer, ESnet, Euclid, Science Gateways
Contact: Linda Vu, +1 510 495 2402, lvu@lbl.gov

Some mysteries of science can only be explained on a nanometer scale —even smaller than a single strand of human DNA, which is about 2.5 nanometers wide. At this scale, scientists can investigate the structure and behavior of proteins that help our bodies fight infectious microbes, and even catch chemical reactions in action. To resolve these very fine details, they rely on synchrotron light sources like the Department of Energy's Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory (Berkeley Lab).

Advanced Light Source (ALS) at Berkeley Lab (Photo by Roy Kaltschmidt)

For decades, synchrotron light sources have been operating on a manual grab-and-go data management model—users travel thousands of miles to run experiments at the football-field-size facilities, download raw data to an external hard drive, then process and analyze the data on their personal computers, often days later. But, a recent deluge of data—brought on by faster detectors and brighter light sources—is quickly making this practice implausible.

CAMERA: Center for Applied Mathematics for Energy Research Applications

CAMERA

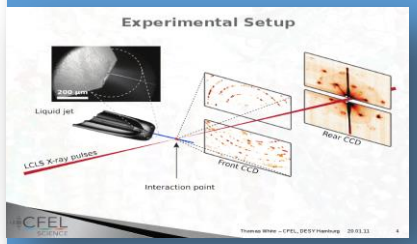
BERKELEY LAB
Lawrence Berkeley National Laboratory

Today:	Tomorrow:	Critical need:	LBNL approach:	New math to:
Facilities data is time-consuming	More data. More quickly. High resolution.	algorithms and analysis for *understanding*	Focused teams of mathematicians/ domain scientists	Guide and optimize experiments

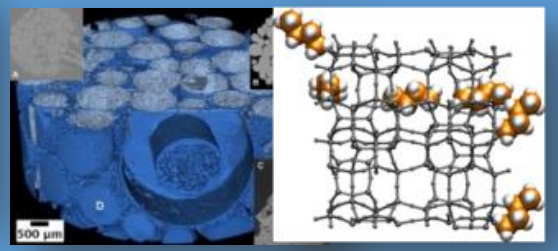
Goal: Build the applied mathematics that helps *transform experimental data into understanding*

Pilot Partners

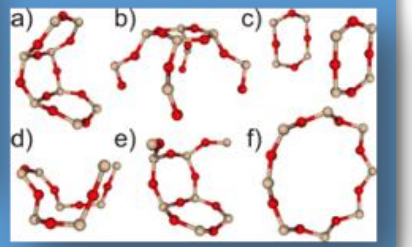
computing *structure* from imaging



analyzing samples and proposed new materials



designing new materials



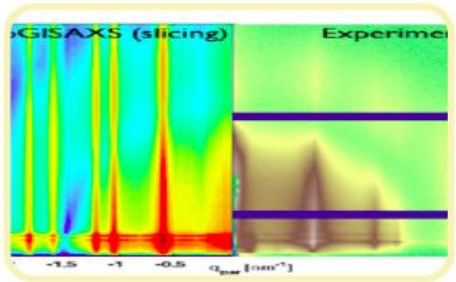
Key: Leverage state-of-the-art mathematics

Spectral clustering
Maximum likelihood estimators
Graph theory
Machine learning
Mori-Zwanzig theory
Computational harmonic analysis
Discrete Galerkin methods
Clique analysis
PDE-based image segmentation
Statistical sampling
Discrete/continuous shape descriptors
Hamilton-Jacobi solvers
Voronoi methods
Representation theory
Bayesian analysis
Optimization methods

Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery

Experimental Facilities

Real-time analysis and Data management



ESnet
ENERGY SCIENCES NETWORK

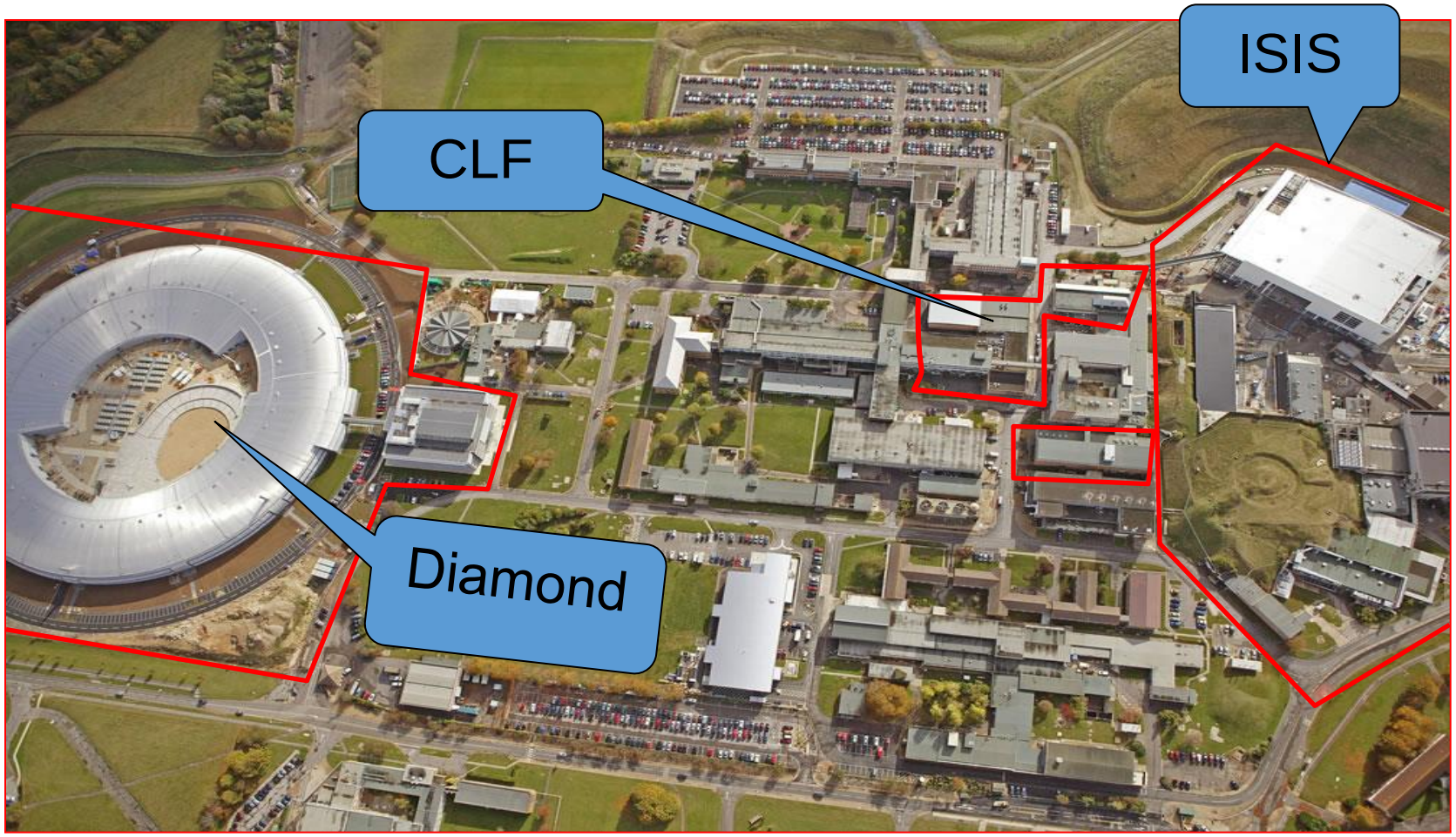
New Mathematical Analyses

Fast Implementations on latest computers

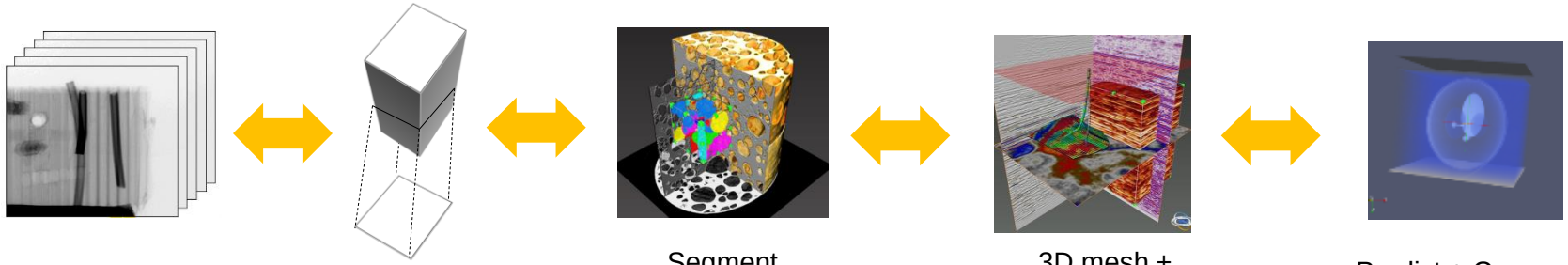
Network for Big Data Science

Computing Facilities

Harwell Site Experimental Facilities



In- and Post-experimental support



Scan | Reconstruct | Segment + Quantify | 3D mesh + Image based Modelling | Predict + Compare

Data Catalogue | Petabyte Data storage | Parallel File system | HPC **CPU+GPU** | Visualisation

Infrastructure + Software + Expertise

ISIS:IMAT

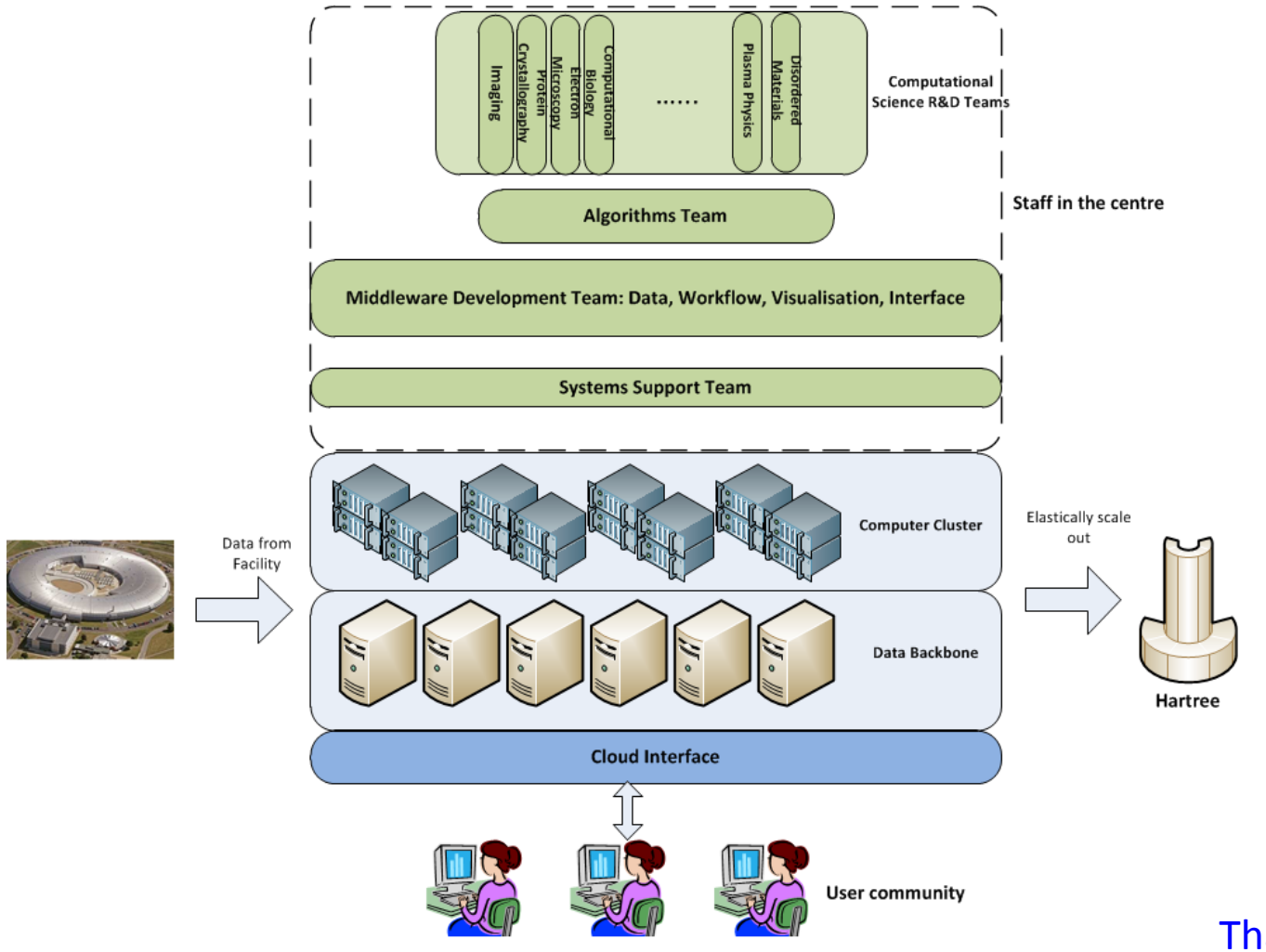
DLS:I12/I13

Erica Yang, Sri Nagella

- **Tomography**: Dealing with high data volumes – 200Gb/scan, ~5 TB/day (one experiment at DLS)
- **MX**: high data volumes, smaller files, but a lot more experiments
- Hard to move the data – needs to be handled at the facility?

Some mage credit: Avizo, Visualization Sciences Group (VSG)

Ada Lovelace Centre Proposal



Thanks to Brian Matthews and Barbara Montanari

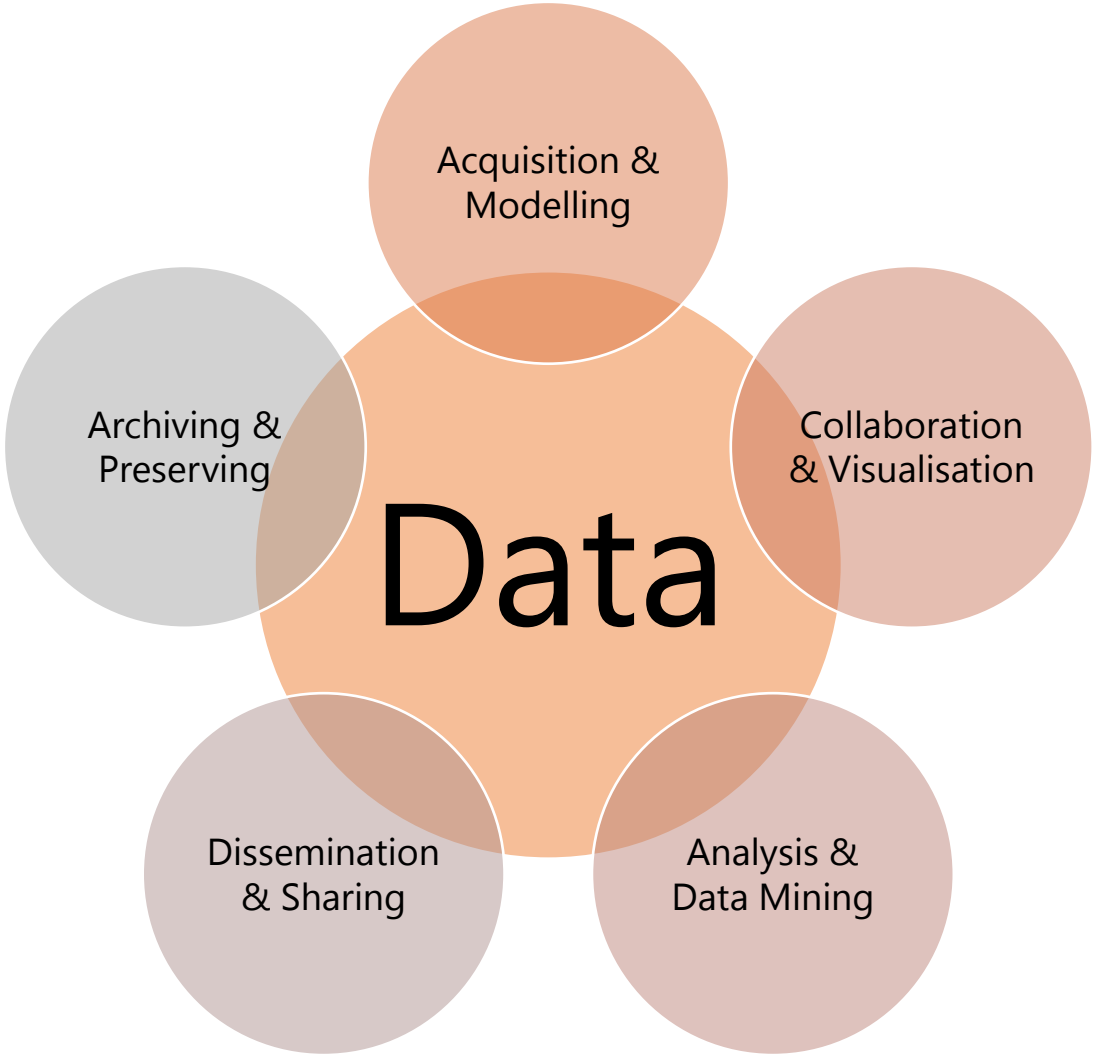
Data Science in the Future?

What is a Data Scientist?

Data Engineer 	**People who are expert at** • Operating at low levels close to the data, write code that manipulates • They may have some machine learning background. • Large companies may have teams of them in-house or they may look to third party specialists to do the work.
Data Analyst 	**People who explore data through statistical and analytical methods** • They may know programming; May be an spreadsheet wizard. • Either way, they can build models based on low-level data. • They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.
Data Steward 	**People who think to managing, curating, and preserving data.** • They are information specialists, archivists, librarians and compliance officers. • This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

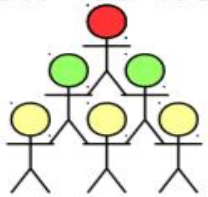
What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda
http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx

The Data-Intensive Research Lifecycle



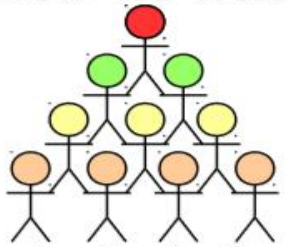
Scientist career paths?

How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

How we work



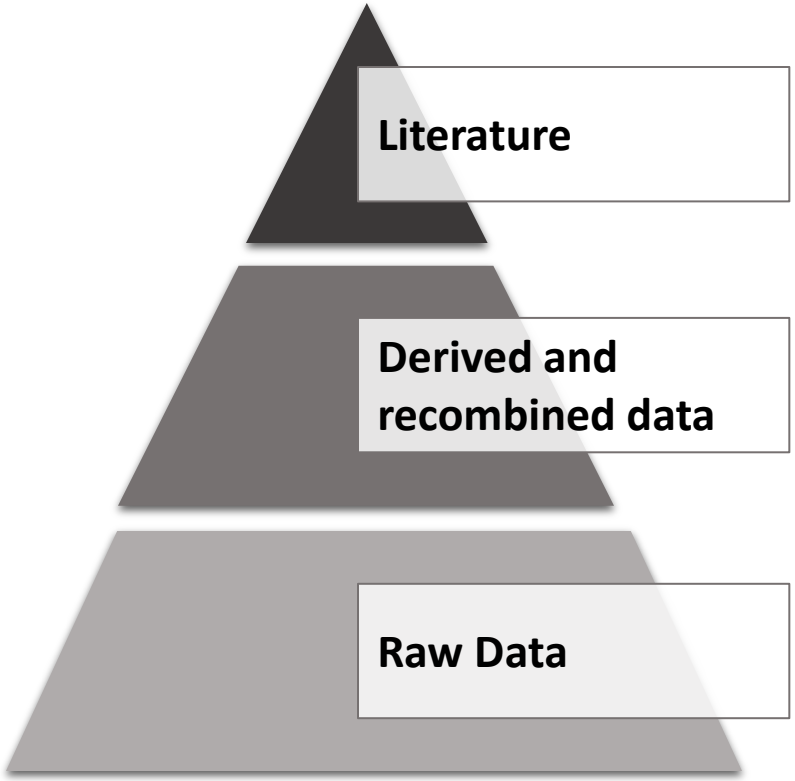
PI stands on the shoulders of her postdocs, students, software engineers and data scientists. (Are the giants down with the turtles?).

- It's fair to say that our institutions have not really caught onto the necessity to have careers for everyone in that stack.
- From the people managing vocabularies and manually entering metadata, to the software engineers and data scientists, we have new careers appearing, and we're not really ready for it.
- Mercifully we're not alone, bioinformatics is blazing a similar trail, but we have much to do.

Slide thanks to Bryan Lawrence

Jim Gray's Vision: All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(*From Jim Gray's last talk)*

Data-Intensive Scientific Discovery