

# **Open Science, Big Data and Research Reproducibility**

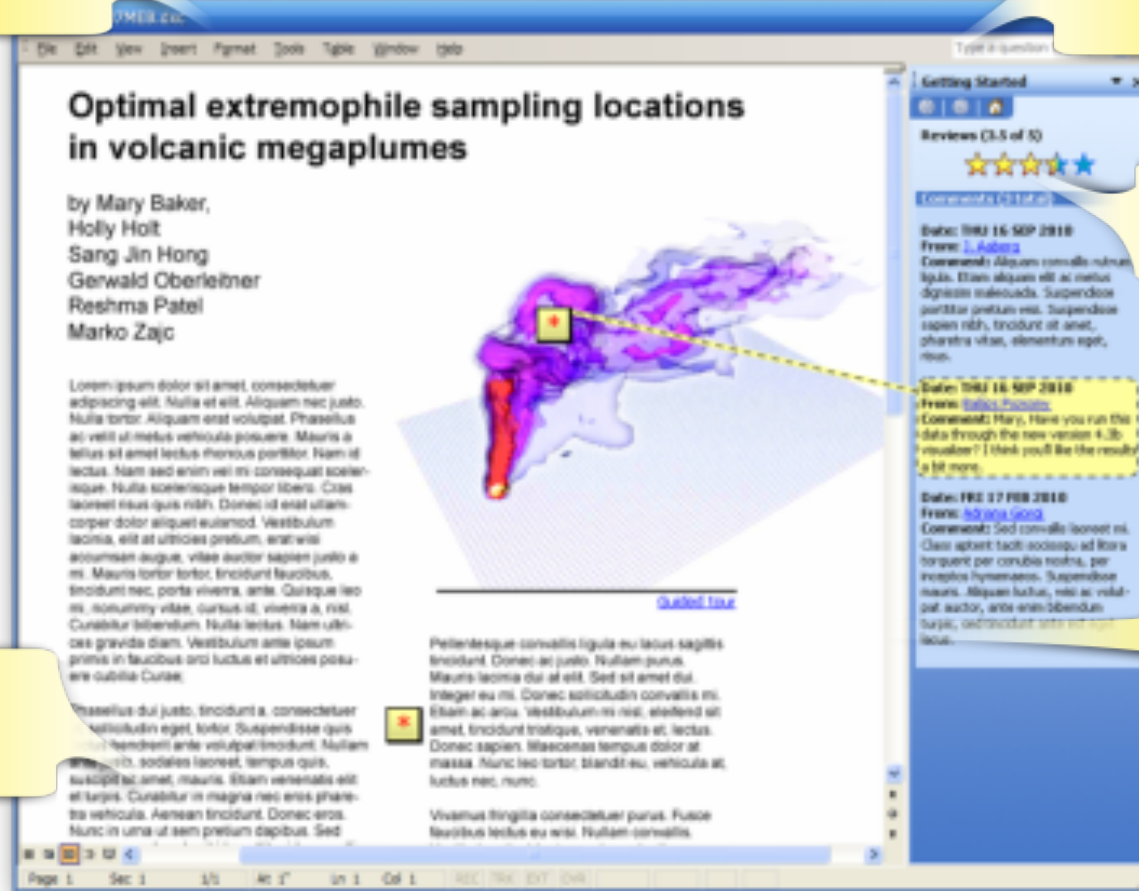
**Tony Hey**  
**Senior Data Science Fellow**  
**eScience Institute**  
**University of Washington**  
[tony.hey@live.com](mailto:tony.hey@live.com)

# **The Vision of Open Science**

# Vision for a New Era of Research Reporting

Reproducible  
Research

Collaboration



Reputation  
& Influence

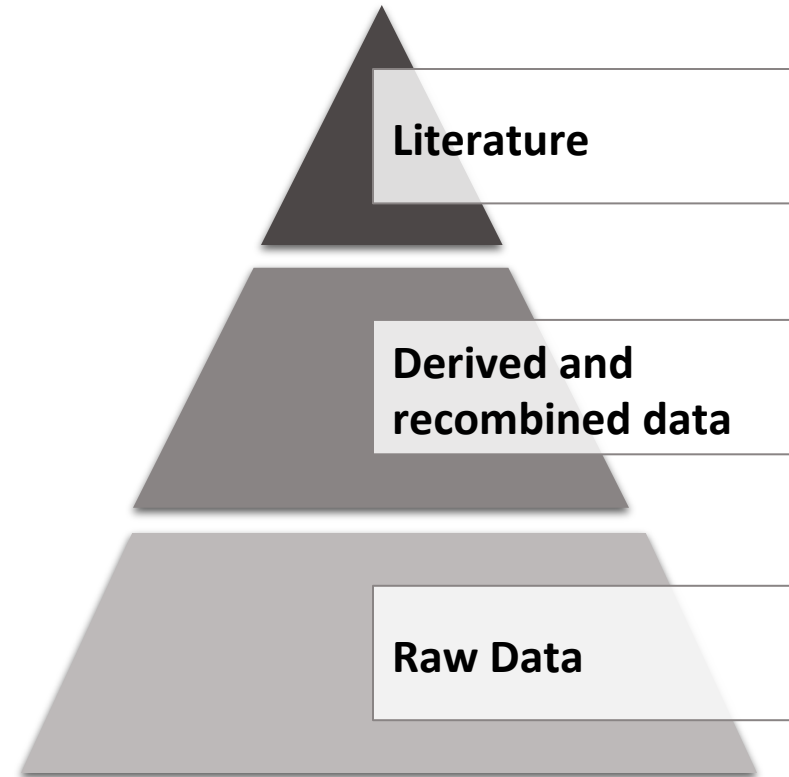
Interactive  
Data

Dynamic  
Documents

*(Thanks to Bill Gates SC05)*

# 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)*

# **OSTP Memo: Open Science and Open Access**

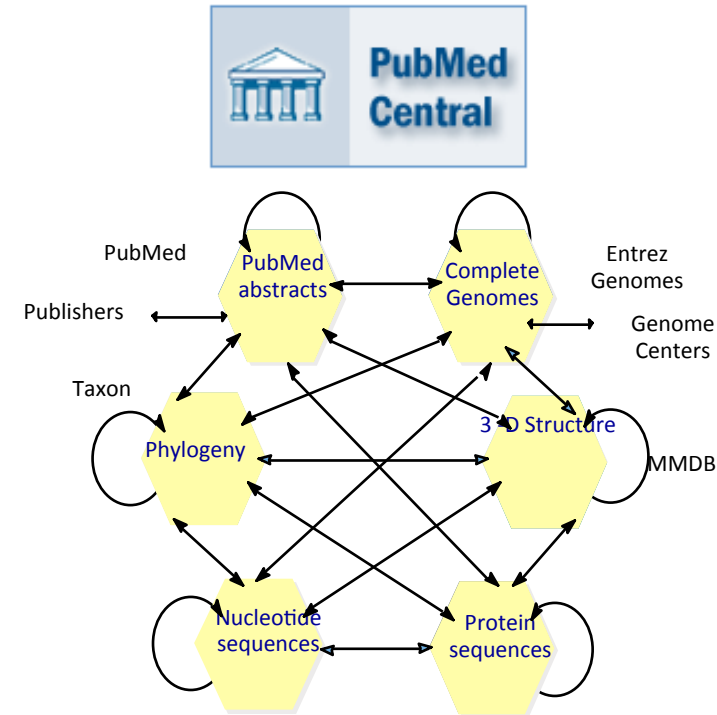
# US OSTP Memorandum

- Directive requiring the major Federal Funding agencies *“to develop a plan to support increased public access to the results of research funded by the Federal Government.”*
- The memorandum defines digital data *“as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens.”*

**22 February 2013**

# 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 tool**

- Jim Gray worked with David Lipman and his team at NCBI to create a 'portable' version of PubMed Central
- This is now deployed in Europe and elsewhere

# US NIH Open Access Policy

- Once posted to PubMed Central, results of NIH-funded research become more prominent, integrated and accessible, making it easier for all scientists to pursue NIH's research priority areas competitively.
- PubMed Central materials are integrated with large NIH research data bases such as Genbank and PubChem, which helps accelerate scientific discovery.
- The policy allows NIH to monitor, mine, and develop its portfolio of taxpayer funded research more effectively, and archive its results “in perpetuity”



# U.S. Department of Energy Increases Access to Results of DOE-funded Scientific Research

August 4, 2014 - 10:49am



## NEWS MEDIA CONTACT

• 202-586-4940

WASHINGTON, D.C. – The U.S. Department of Energy is introducing new measures to increase access to scholarly publications and digital data resulting from Department-funded research.

The Energy Department has launched the Public Access Gateway for Energy and Science – **PAGES** – a web-based portal that will provide free public access to accepted peer-reviewed manuscripts or published scientific journal articles within 12 months of publication.

“Increasing access to the results of research funded by the Department of Energy will enable researchers and entrepreneurs to capitalize on our substantial research and development investments,” said Secretary of Energy Ernest Moniz. “These new policies set the stage for increased innovation, commercial opportunities, and accelerated scientific breakthroughs.”

As it grows in content, PAGES will include access to DOE-funded authors’ accepted manuscripts hosted primarily by the Energy Department’s National Labs and grantee institutions, in addition to the public access offerings of publishers. For publisher-hosted content, the Department is collaborating with the publisher consortium CHORUS -- the Clearinghouse for the Open Research of the United States.

## RELATED ARTICLES

[Secretary Abraham Announces Energy Department “What’s Next” Expo to be Held in Detroit Area](#)

[Access to Science Information Expands with Science.gov 5.0 Launch](#)



[Digital Strategy](#)



Speeding access to science information from DOE and beyond



HOME

ABOUT OSTI

SCIENCE SEARCH  
TOOLS

DOE PAGES<sup>Beta</sup> /  
PUBLIC ACCESS

COMMUNICATIONS

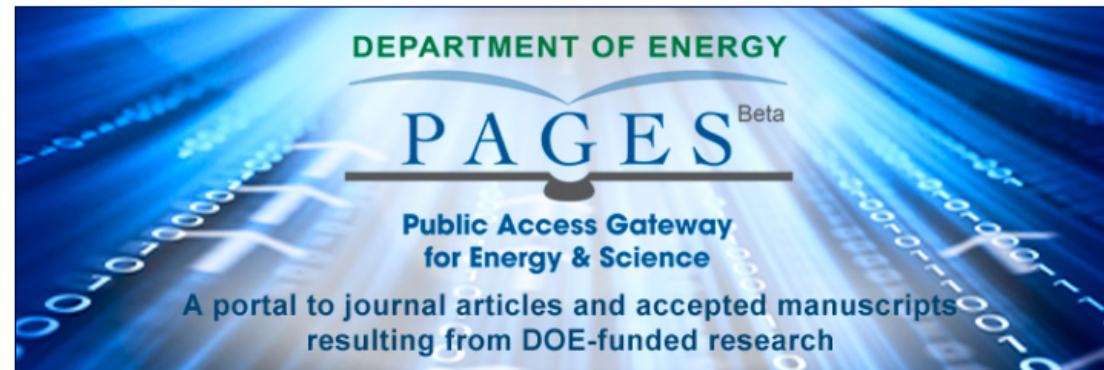
DOE STI PROGRAM

Find DOE R&D Results

GO

SciTech Connect

**DOE Scientific and Technical Information...and more**



For additional information, see the OSTI Catalogue of Collections.

**News/ Blogs**

*Accepted Manuscript Submissions for DOE PAGES<sup>Beta</sup> Officially Start October 1, 2014*

*U.S. Department of Energy Increases Access to Results of DOE-funded Scientific Research*

*Achieving Public Access: The Department of Energy Launches DOE PAGES<sup>Beta</sup>*

*Brian Hitson Named Director of DOE OSTI*

# NIH Open Access Compliance?

- PMC Compliance Rate

- Before legal mandate compliance was 19%
- Signed into law by George W. Bush in 2007
- After legal mandate compliance up to 75%

- NIH have taken a further step of announcing that, 'sometime in 2013' they stated that they

*'... will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.'*

- NIH now implemented their policy about continuation awards

- Compliance rate increasing ½% per month
- By November 2014, compliance rate had reached 86%

# Open Access to Scholarly Publications and Data: 2013 as the Tipping Point?

- US OSTP Memorandum 26 February 2013
- Global Research Council Action Plan 30 May 2013
- G8 Science Ministers Joint Statement 12 June 2013
- European Union Parliament 13 June 2013

# University of California approves Open Access

- UC is the largest public research university in the world and its faculty members receive roughly 8% of all research funding in the U.S.
- UC produces 40,000 publications per annum corresponding to about 2 – 3 % of all peer-reviewed articles in world each year
- UC policy requires all 8000 faculty to deposit full text copies of their research papers in the UC eScholarship repository unless they specifically choose to opt-out

2 August 2013

# Data Curation

# PUBLIC ACCESS

## To Results of NSF-funded Research

[Public Access Home](#)[Plan](#)[Executive Summary](#)[Press Release](#)[Frequently Asked Questions \(FAQs\)](#)[Search NSF Awards](#)[NSF Public Access Feedback](#)

The National Science Foundation (NSF or Foundation) has developed a plan outlining a framework for activities to increase public access to scientific publications and digital scientific data resulting from research the foundation funds. The plan, entitled "Today's Data, Tomorrow's Discoveries," is consistent with the objectives set forth in the Office of Science and Technology Policy's Feb. 22, 2013, memorandum, "Increasing Access to the Results of Federally Funded Research," and with long-standing policies encouraging data sharing and communication of research results.

As outlined in section 3.1 of the plan, NSF will require that either the version of record or the final accepted manuscript in peer-reviewed scholarly journals and papers in juried conference proceedings or transactions must:

- Be deposited in a public access compliant repository designated by NSF;
- Be available for download, reading and analysis free of charge no later than 12 months after initial publication;
- Possess a minimum set of machine-readable metadata elements in a metadata record to be made available free of charge upon initial publication;
- Be managed to ensure long-term preservation; and
- Be reported in annual and final reports during the period of the award with a persistent identifier that provides links to the full text of the publication as well as other metadata elements.

This NSF requirement will apply to new awards resulting from proposals submitted, or due, on or after the effective date of the *Proposal & Award Policies & Procedures Guide (PAPPG)* that will be issued in **January 2016**.

# New Requirements for DOE Research Data

The Energy Department's Office of Science also has issued new requirements regarding management of digital research data by Office of Science-supported researchers. All proposals for research funding submitted to the Office of Science will be required to include a Data Management Plan that describes whether and how the digital research data generated in the course of the proposed research will be shared and preserved.

The new requirements regarding management of digital research data will appear in funding solicitations and invitations issued by the Office of Science beginning Oct. 1, 2014. A statement of the new requirements, including guidance on the development of a Data Management Plan, can be found on the [Office of Science website](#). Other Energy Department research offices will implement data management plan requirements within the next year.



# EPSRC Expectations for Data Preservation

- Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires
- Research organisations will ensure that effective data curation is provided throughout the full data lifecycle, with 'data curation' and 'data lifecycle' being as defined by the Digital Curation Centre

# Progress in Data Curation?

Professor James Frew (UCSB):

- Biggest change is funding agency mandate.
  - NSF's Data Management Plan for all proposals has made scientists (pretend?) to take data curation seriously.
  - There are better curated databases and metadata now - but not sure that quality fraction is increasing!
  - Frew's first law: scientists don't write metadata
  - Frew's second law: any scientist can be forced to write bad metadata
- 
- Should automate creation of metadata as far as possible
  - Scientists need to work with metadata specialists with domain knowledge

# **Open Science and Research Reproducibility**

# Jon Claerbout and the Stanford Exploration Project (SEP) with the oil and gas industry

- Jon Claerbout is the Cecil Green Professor Emeritus of Geophysics at Stanford University
- He was one of the first scientists to recognize that the reproducibility of his geophysics research required access not only to the text of the paper but also to the data being analyzed and the software used to do the analysis
- His 1992 Paper introduced an early version of an 'executable paper'

## Electronic Documents Give Reproducible Research a New Meaning

*Jon Claerbout and Martin Karrenbach*

*This was an invited paper at the October 25-29, 1992 meeting of the Society of Exploration Geophysics and it appears in the program as this extended abstract.*

### ABSTRACT

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a concrete definition of reproducibility in computationally oriented research. Experience at the Stanford Exploration Project shows that preparing such electronic documents is little effort beyond our customary report writing; mainly, we need to file everything in a systematic way.

# Preface to SEP report 124

2/22/2006

The electronic version of this report <http://sepwww.stanford.edu/private/docs/sep124> makes the included programs and applications available to the reader. The markings [ER], [CR], and [NR] are promises by the author about the reproducibility of each figure result. Reproducibility is a way of organizing computational research that allows both the author and the reader of a publication to verify the reported results. Reproducibility facilitates the transfer of knowledge within SEP and between SEP and its sponsors.

ER

denotes Easily Reproducible and are the results of processing described in the paper. The author claims that you can reproduce such a figure from the programs, parameters, and makefiles included in the electronic document. The data must either be included in the electronic distribution, be easily available to all researchers (e.g., SEG-EAGE data sets), or be available in the SEP data library <http://sepwww.stanford.edu/public/docs/sepdatilib/toc.html>.

We assume you have a UNIX workstation with Fortran, Fortran90, C, X-Windows system and the software downloadable from our website (SEP makerules, SEPlib, and the SEP latex package), or other free software such as SU. Before the publication of the electronic document, someone other than the author tests the author's claim by destroying and rebuilding all ER figures. Some ER figures may not be reproducible by outsiders because they depend on data sets that are too large to distribute, or data that we do not have permission to redistribute but are in the SEP data library.

CR

denotes Conditional Reproducibility. The author certifies that the commands are in place to reproduce the figure if certain resources are available. SEP staff have only attempted to make sure that the makefile rules exist and the source codes referenced are provided. The primary reasons for the CR designation is that the processing requires 20 minutes or more, or commercial packages such as Matlab or Mathematica.

M

denotes a figure that may be viewed as a movie in the web version of the report. A movie may be either ER or CR.

NR

denotes Non-Reproducible figures. SEP discourages authors from flagging their figures as NR except for figures that are used solely for motivation, comparison, or illustration of the theory, such as: artist drawings, scannings, or figures taken from SEP reports not by the authors or from non-SEP publications.

Our testing is currently limited to LINUX 2.4 (using the Portland Group Fortran90 compiler), but the code should be portable to other architectures. Reader's suggestions are welcome. For more information on reproducing SEP's electronic documents, please visit <http://sepwww.stanford.edu/research/redoc/>.

# Serious problems of research reproducibility in bioinformatics

- Review of 2,047 retracted articles indexed in PubMed in May of 2012 concluded that:
  - 21.3% were retracted because of errors,
  - 67.4% were retracted because of scientific misconduct
    - Fraud or suspected fraud (43.4%)
    - Duplicate publication (14.2%)
    - Plagiarism (9.8%)
- Study by pharma companies Bayer and Amgen concluded that between 60% and 70% of biomedicine studies may be non-reproducible
  - Amgen scientists were only able to reproduce 7 out of 53 cancer results published in Science and Nature

## Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at [go.nature.com/huhbyr](http://go.nature.com/huhbyr)). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the *Nature* research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility ([go.nature.com/oloqip](http://go.nature.com/oloqip)). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at *Nature* journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters that can introduce bias or influence robustness, and provide precise characterization of key reagents that may be subject to biological variability, such as cell lines and antibodies. The checklist also consolidates existing policies about data deposition and presentation.

We will also demand more precise descriptions of statistics, and

we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange ([www.nature.com/protocolexchange](http://www.nature.com/protocolexchange)), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require the commitment of funders, institutions, researchers and publishers. It is encouraging that NIH institutes have led community discussions on this topic and are considering their own recommendations. We urge others to take note of these and of our initiatives, and do whatever they can to improve research reproducibility. ■

Nature  
+ other  
publishers

Allow  
details  
of methods  
and provide  
data

# Computational Science and Reproducibility



# eScience 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**

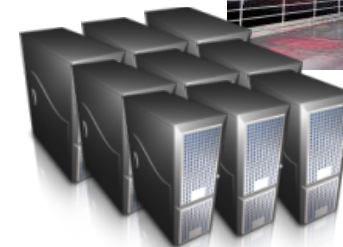
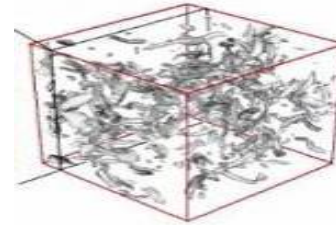
- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

**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



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



*(With thanks to Jim Gray)*

# 2012 ICERM Workshop on Reproducibility in Computational and Experimental Mathematics

- The workshop participants noted that computational science poses a challenge to the usual notions of ‘research reproducibility’
- Experimental scientists are taught to maintain lab books that contain details of the experimental design, procedures, equipment, raw data, processing and analysis (but ...)
- Few computational experiments are documented so carefully:
  - Typically there is no record of the workflow, no listing of the software used to generate the data, and inadequate details of the computer hardware the code ran on, the parameter settings and any compiler flags that were set

# Best Practices for Researchers Publishing Computational Results

- **Data must be available and accessible.** In this context the term "data" means the raw data files used as a basis for the computations, that are necessary for others to regenerate published computational findings.
- **Code and methods must be available and accessible.** The traditional methods section in a typical publication does not communicate sufficient detail for a knowledgeable reader to replicate computational results. A necessary action is making the complete set of instructions, typically in the form of computer scripts or workflow pipelines, conveniently available.
- **Citation.** Do it. If you use data you did not collect from scratch, or code you did not write, however little, cite it. Citation standards for code and data are discussed but it is less important to get the citation perfect than it is to make sure the work is cited at all.
- **Copyright and Publisher Agreements.** Publishers, almost uniformly, request that authors transfer all ownership rights over the article to them. All they really need is the authors' permission to publish.
- **Supplemental materials.** Publishers should establish style guides for supplemental sections, and authors should organize their supplemental materials following best practices.

From <http://wiki.stodden.net>

# But Sustainability of Data Links?

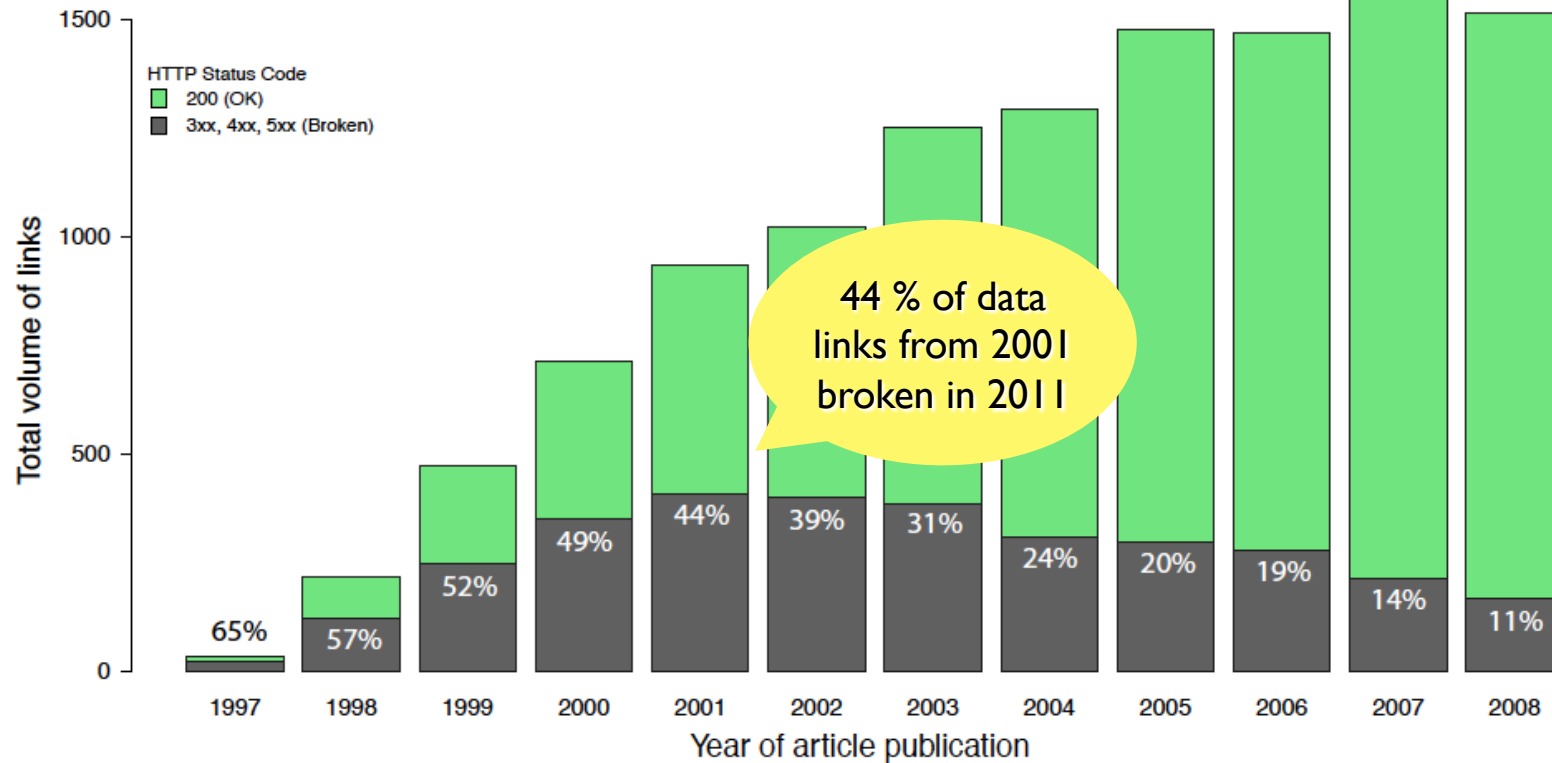


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

# Challenge of Numerical Reproducibility?

‘Numerical round-off error and numerical differences are greatly magnified as computing simulations are scaled up to run on highly parallel systems. As a result, it is increasingly difficult to determine whether a code has been correctly ported to a new system, because computational results quickly diverge from standard benchmark cases. And it is doubly difficult for other researchers, using independently written codes and distinct computer systems, to reproduce published results.’

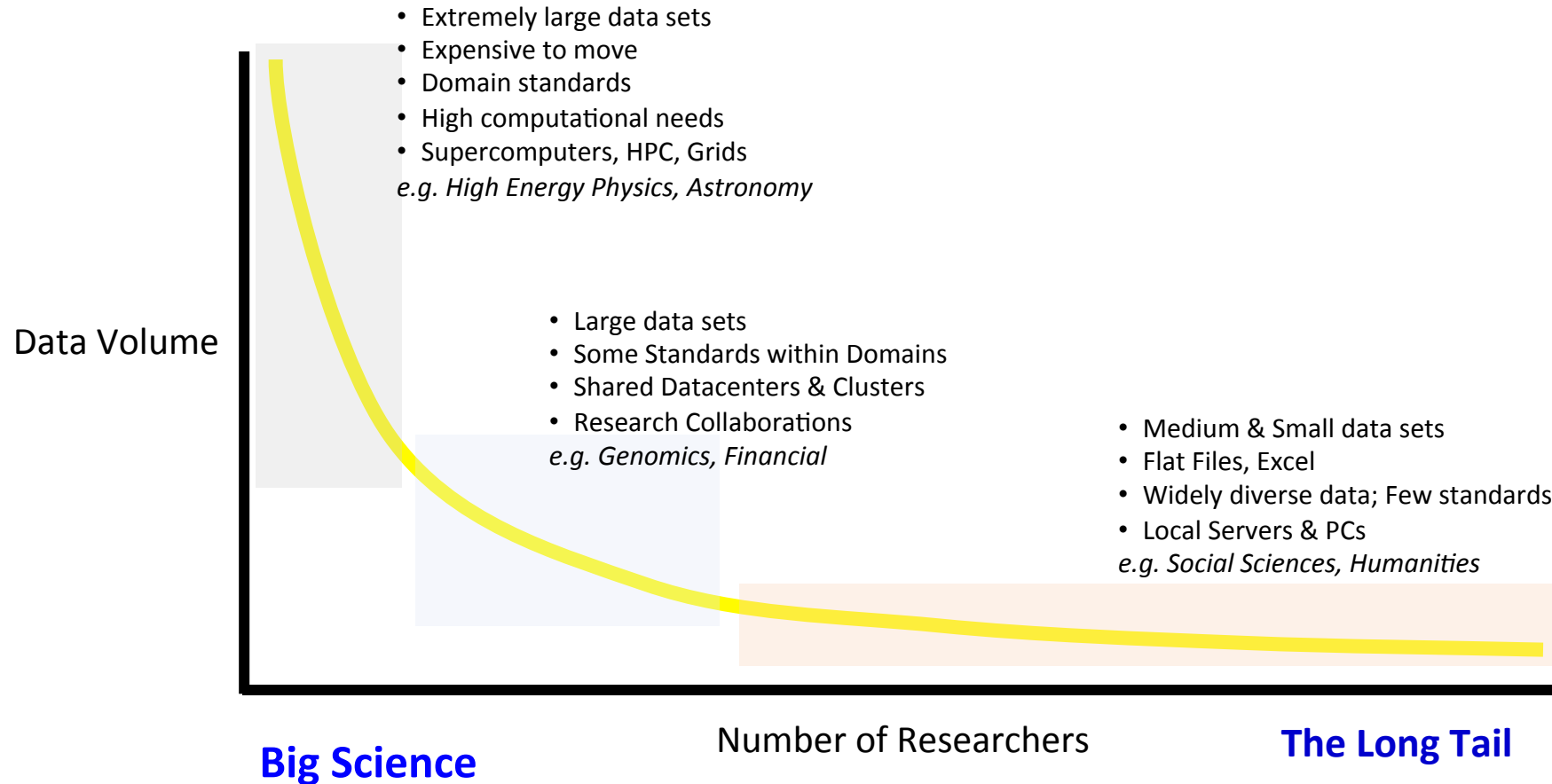
David Bailey: ‘Fooling the Masses: Reproducibility in High-Performance Computing’ (<http://www.davidhbailey.com>)

# Same Physics, Different Programs?

- Different programs written by different researchers can be used to explore the physics of the same complex system
  - Programs may use different algorithms and/or different numerical methods
  - Codes are different but the target physics problem is the same
  - Cannot insist on exact numerical agreement
- Computational reproducibility involves finding 'similar' quantitative results for the key physical parameters of the system being explored

# **Big Science and the Long Tail**

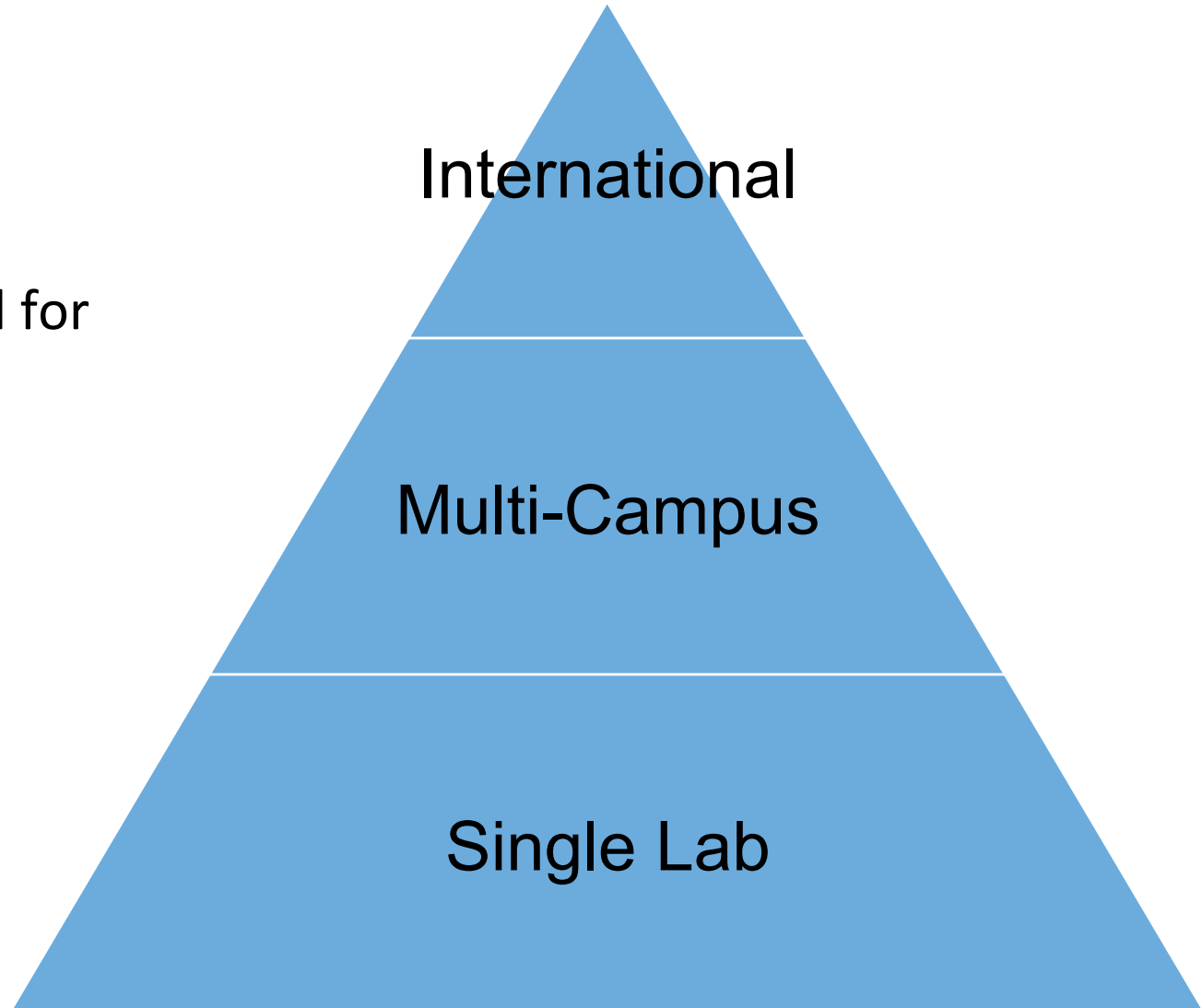
# Big Science and the Long Tail





# Project Pyramids

- In most disciplines there are:
  - a few “giga” projects,
  - several “mega” consortia
  - and then many small labs.
- Often some instrument creates need for giga-or mega-project:
  - Polar station
  - Accelerator
  - Telescope
  - Remote sensor
  - Genome sequencer
  - Supercomputer
- Tier 1, 2, 3 facilities to use instrument + data



# Experiment Budgets $\frac{1}{4}$ ... $\frac{1}{2}$ Software

## Software for

- Instrument scheduling
- Instrument control
- Data gathering
- Data reduction
- Database
- Analysis
- Modeling
- Visualization

## Millions of lines of code

Repeated for experiment after experiment

Not much sharing or learning

## CS can change this

## Build generic tools

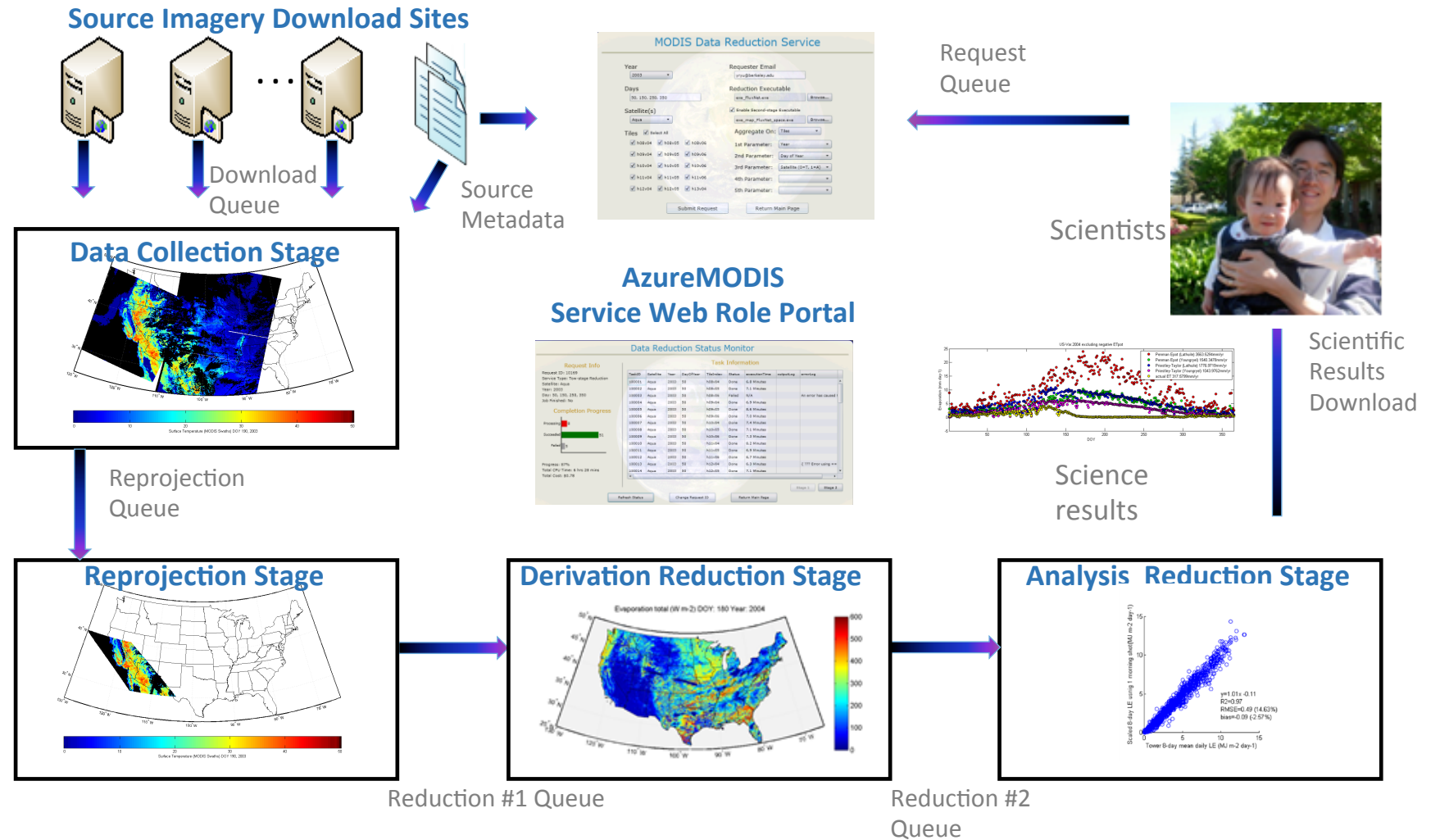
- Workflow schedulers
- Databases and libraries
- Analysis packages
- Visualizers
- ...

# Computing and Big Science

- Computational science – in the sense of performing computer simulations of complex systems – is only one aspect of computing in physics
- For Big Science projects, generation and analysis of the data would not be possible without large-scale computing resources
- Examples:
  - NASA MODIS Satellite Data Pre-processing
  - LSST Large Synoptic Survey Telescope
  - LHC Large Hadron Collider Experiments
- New challenges for open science

# NASA MODIS Satellite: Image Processing Pipeline

- Data collection stage
  - Downloads requested input tiles from NASA ftp sites
  - Includes geospatial lookup for non-sinusoidal tiles that will contribute to a reprojected sinusoidal tile
- Reprojection stage
  - Converts source tile(s) to intermediate result sinusoidal tiles
  - Simple nearest neighbor or spline algorithms
- Derivation reduction stage
  - First stage visible to scientist
  - Computes ET in our initial use
- Analysis reduction stage
  - Optional second stage visible to scientist
  - Enables production of science analysis artifacts such as maps, tables, virtual sensors



<http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx>

Slide thanks to Catharine van Ingen

# LSST Data Management

Building a high-performance, scalable, general purpose, open source O/R survey data processing and analysis system.

A subsystem of the [LSST Project](#).



## Building the Next-Generation Data Processing System

The [LSST](#) is a large optical survey project funded by the National Science Foundation and the Department of Energy. It will continually image the sky, identify changes in near real time, and over a decade of operations collect tens of petabytes of data building up the deepest, widest, image of the Universe. Its data will enable a range of science goals from identification of Near Earth Asteroids to understanding the nature of Dark Energy.

A survey of this scale requires significant computing resources but also a modern, high-performance, scalable, data processing and analysis system. The LSST Data Management team is guiding an effort to build such a suite. Primarily written in Python and C++, open source, and comprised of modular codes ranging from science pipelines to web user interfaces, the LSST software stack will power the LSST and form a basis that other projects can reuse in the future.

The LSST DM team is distributed across a number of partner institutions — the [LSST Project Office](#), the [Infrared Processing and Analysis Center](#), the [National Center for Supercomputing Applications](#), [Princeton University](#), [SLAC National Accelerator Laboratory](#), and the [University of Washington](#) — but also helped by contributors from the community, the LSST science collaborations, and other project subsystems.



[Learn more about LSST data processing »](#)

### Science Pipelines

The [LSST Science Pipelines](#) will implement the core image processing and data analysis algorithms needed to process optical survey imaging data at low latency and unprecedented scale and accuracy. We are writing pipelines for single-epoch image processing, coaddition, image differencing, optimal multi-epoch measurements, and (global) photometric and astrometric calibration, among others.

### Scalable Database

To satisfy the need to efficiently store, query, and analyze catalogs running into trillions of rows and petabytes of data, we are developing [Qser](#), a distributed shared-nothing SQL database query system.

### User Interface

One of the most important jobs of a large survey is to provide access. This includes access to catalogs, processed images, and raw images. Access in the next generation of surveys will extend to visualization and analysis. We are writing interfaces that will allow thousands of users to query, download, visualize, and analyze petabytes of LSST data.

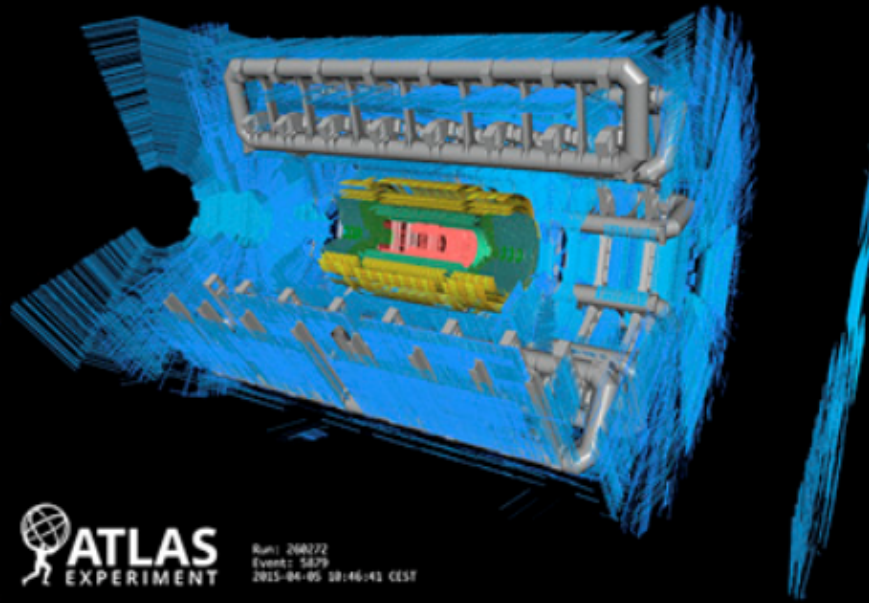
### Data Access Middleware

In order to build a scalable, portable processing system, we are creating extensible middleware to transparently access data irrespective of storage location or format.

### Distributed Execution

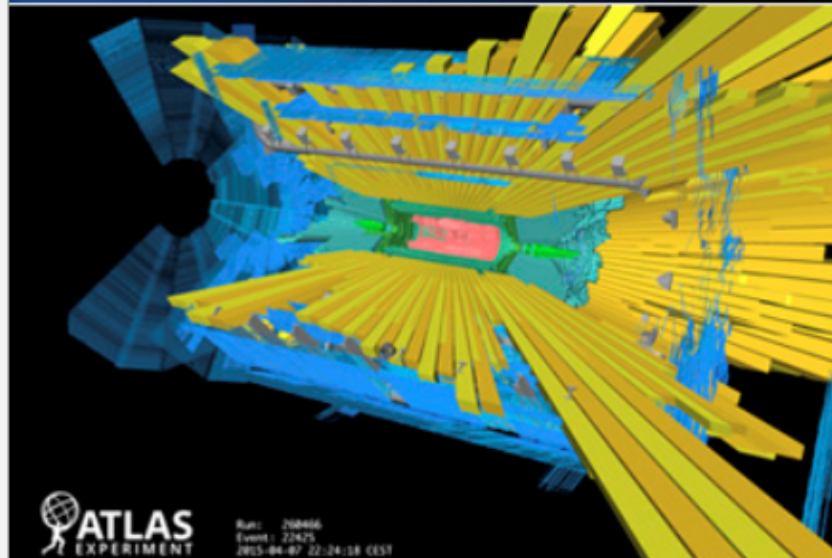
The LSST data processing pipelines will need to efficiently scale from single core execution to tens of thousands of cores. To meet this requirement we are building an orchestration framework to launch and monitor jobs on many different systems at many different scales.

## LHC and ATLAS Restart



ATLAS Is Ready and Waiting for Collisions

## ATLAS News



### Splashes for Synchronization

ATLAS uses "beam splash" events to provide simultaneous signals to large parts of the detector, and verify that the readout of different detectors elements are fully synchronized. [More...](#)

# The ATLAS Software: Thanks to Gordon Watts

- Software written by around 700 postdocs and grad students
- ATLAS software is 6M lines of code – 4.5M in C++ and 1.5M in Python
- Typical reconstruction task has 400 to 500 software modules
- Software system begins with data acquisition of collision events from 100M readout channels and then reconstructs particle trajectories
- The reconstruction process requires a detailed Monte Carlo simulation of the ATLAS detector taking account of the geometries, properties and efficiencies of each subsystem of the detector
- Produces values for the energy and momentum of the tracks observed in the detector
- Then find Higgs boson 😊

## Fact Sheet 1

[Download print version \(PDF\)](#)

### The ATLAS Detector

Diameter: 25m

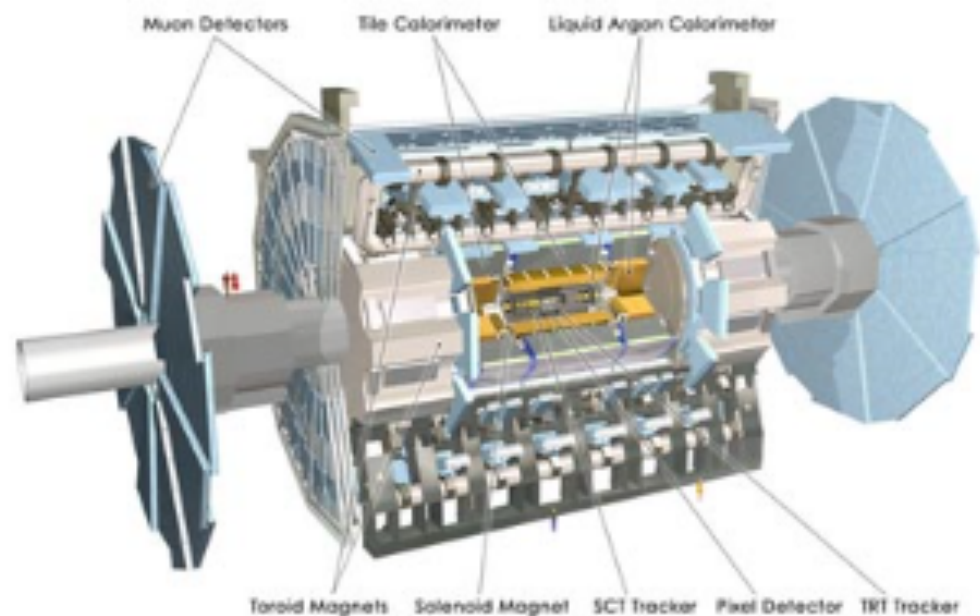
Length: 46m

Barrel Toroid Length: 26m

Overall weight: 7000 tonnes

~100 million electronic channels

3000 km of cables





## Calorimeters

Measure the energies carried by the particles

### Liquid Argon (LAr) Calorimeter

- Barrel 6.4 m long, 53 cm thick, 110,000 channels.
- Works with Liquid Argon at  $-183^{\circ}\text{C}$
- LAr endcap consists of the forward calorimeter, electromagnetic (EM) and hadronic endcaps.
- EM endcaps each have thickness 0.632 m and radius 2.077 m.
- Hadronic endcaps consist of two wheels of thickness 0.8 m and 1.0 m with radius 2.09 m.
- Forward calorimeter has three modules of radius 0.455 m and thickness 0.450 m each.

### Tile Calorimeter (TileCal)

- Barrel made of 64 wedges, each 5.6 m long and 20 tonnes.
- Each Endcap has 64 wedges, each 2.8 m long.
- 500,000 plastic scintillator tiles.

## Muon System

Identifies and measures the momenta of muons

### Thin Gap Chambers

For triggering and 2nd coordinate measurement (non-bending direction) at ends of detector.

- 440,000 channels

### Resistive Plate Chambers

For triggering and 2nd coordinate measurement in central region.

- 380,000 channels
- Electric Field 5,000 V/mm

### Monitored Drift Tubes

Measure curves of tracks.

- 1,171 chambers with total 354,240 tubes (3 cm diameter, 0.85-6.5 m long).
- Tube resolution  $80\ \mu\text{m}$

### Cathode Strip Chambers

Measure precision coordinates at ends of detector.

- 70,000 channels
- Resolution  $60\ \mu\text{m}$

# ATLAS Software Engineering Methodologies

- Automated integration testing of modules
- Candidate release code versions tested in depth by running long jobs, producing 'standard' plots, and detailed comparison with reference data sets
- ATLAS uses JIRA tool for bug tracking
- After every observed difference has been investigated and resolved, the new version of the code is released to whole ATLAS collaboration
- ATLAS uses Apache Subversion (SVN) version-control system.
- With over 2000 software packages to be tracked, ATLAS uses release management software developed by the collaboration

# Research Reproducibility at the LHC?

- At the LHC there are the two experiments - ATLAS and CMS - looking for new 'Higgs and beyond' physics
- The detectors and the software used by these two experiments are very different
- The two experiments are at different intersection points of the LHC and generate different data sets
- Research reproducibility is addressed by having the same physics observed in different experiments: e.g. see the Higgs boson at the same mass value in both experiments
- Making meaningful data available to the public is difficult but the new CERN Open Data portal is now making a start ...

# Education



The CMS (Compact Muon Solenoid) experiment is one of two large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics such as the characteristics of the Higgs boson, extra dimensions or dark matter.

Explore CMS >



ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms. More than 1000 scientists are part of the

Explore ALICE >



The ATLAS (A Toroidal LHC ApparatuS) experiment is a general purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces, and evidence for dark matter candidates in the Universe.

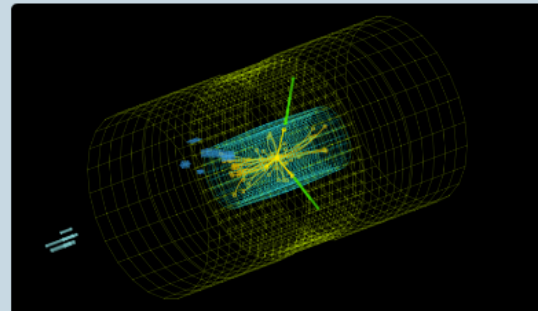
Explore ATLAS >



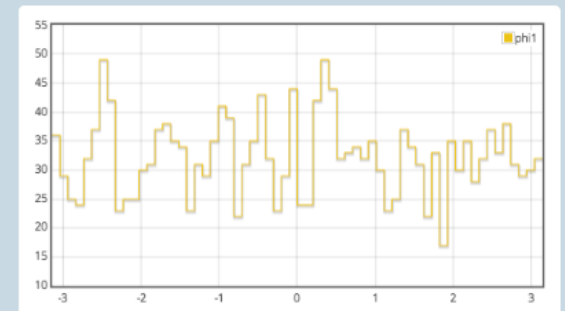
The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

Explore LHCb >

For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here



Visualise events >



Visualise histograms >

Learning Resources >

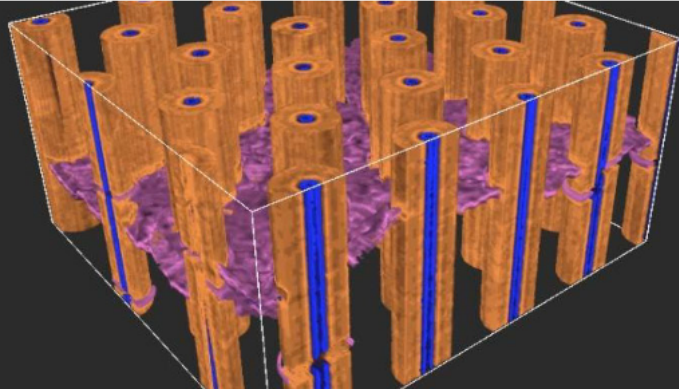
<http://opendata.cern.ch>

# Medium-Scale Science: An Example from the UK

- In UK, the Research Funding Agency RCUK (approx. NSF equivalent) supports a dozen or so 'Collaborative Computational Projects' (CCPs)
- Each CCP is focused on a small number of scientific codes and provides support to a specific community
- Team of scientific software experts at Daresbury Lab helps university researchers develop and maintain codes

UK Collaborative Computational Projects

Home About CCPs Current CCPs Former CCPs Steering Panel Contact



Current CCPs

Please follow the links in the table to visit the individual CCP web pages.

| CCP                         | Chair                  | Title   |
|-----------------------------|------------------------|---|
| <a href="#">CCP4</a>        | Prof David Brown       | Macromolecular Crystallography  |
| <a href="#">CCP5</a>        | Prof Stephen Parker    | The Computer Simulation of Condensed Phases                               |
| <a href="#">CCP9</a>        | Prof Mike Payne        | Computational Electronic Structure of Condensed Matter                    |
| <a href="#">CCP12</a>       | Prof Stewart Cant      | High Performance Computing in Engineering                                 |
| <a href="#">CCP-ASEArch</a> | Prof Mike Giles        | Algorithms and Software for Emerging Architectures                        |
| <a href="#">CCP-BioSim</a>  | Prof Adrian Mulholland | Biomolecular Simulation at the Life Sciences Interface                    |
| <a href="#">CCP-EM</a>      | Dr Martyn Winn         | Electron Cryo-Microscopy  |
| <a href="#">CCP1</a>        | Prof Phillip Withers   | Tomographic Imaging   |
| <a href="#">CCPN</a>        | Prof Geerten Vuister   | NMR   |
| <a href="#">CCP-NC</a>      | Dr Jonathan Yates      | NMR Crystallography   |
| <a href="#">CCP-Plasma</a>  | Dr Tony Arber          | Computational Plasma Physics  |
| <a href="#">CCPQ *</a>      | Prof Tania Monteiro    | Quantum Dynamics in Atomic, Molecular and Optical Physics                 |
| <a href="#">CCP-SAS</a>     | Prof Steve Perkins     | Analysis of Structural Data in Chemical Biology and Soft Condensed Matter |
| <a href="#">CCPForge</a>    | Prof Chris Greenough   | Collaborative Software Development Environment Tool                       |

\* CCPQ was formed from [CCP2](#) "Continuum States of Atoms and Molecules", incorporating aspects of [CCP6](#) "Molecular Quantum Dynamics".

EPSC MRC Science & Technology Facilities Council BBSRC cecam

# Long Tail Science: Small research groups

- Typically faculty professor and a few postdocs and graduate students
- The researchers usually have little or no formal training in software development and software engineering technologies
- Fortran is still used by many 'Long Tail' scientists although there is increasing use of Python
- Small research groups have no software budget so use packages like MATLAB or FLUENT
- Excel is used extensively for data analysis and management and visualization

# **Software engineering and software carpentry**

# How do scientists develop and use scientific software?

 Full Text  
Sign-In or Purchase

Need Full-Text?

Request a free trial to IEEE Xplore for your organization.

FREE TRIAL

6  
Author(s)

Hannay, J.E. ; Dept. of Software Eng., Univ. of Oslo, Oslo ; Langtangen, H.P. ; MacLeod, C. ; Pfahl, D.  
[more authors](#)

Abstract

Authors

References

Cited By

Keywords

Metrics

Similar

 Download Citations

 Email

 Print

 Request Permissions

 Save to Project



New knowledge in science and engineering relies increasingly on results produced by scientific software. Therefore, knowing how scientists develop and use software in their research is critical to assessing the necessity for improving current development practices and to making decisions about the future allocation of resources. To that end, this paper presents the results of a survey conducted online in October-December 2008 which received almost 2000 responses. Our main conclusions are that (1) the knowledge required to develop and use scientific software is primarily acquired from peers and through self-study, rather than from formal education and training; (2) the number of scientists using supercomputers is small compared to the number using desktop or intermediate computers; (3) most scientists rely primarily on software with a large user base; (4) while many scientists believe that software testing is important, a smaller number believe they have sufficient understanding about testing concepts; and (5) that there is a tendency for scientists to rank standard software engineering concepts higher if they work in large software development projects and teams, but that there is no uniform trend of association between rank of importance of software engineering concepts and project/team size.

**Published in:**

[Software Engineering for Computational Science and Engineering, 2009. SECSE '09. ICSE Workshop on](#)

**Date of Conference:**

23-23 May 2009

IEEE Access

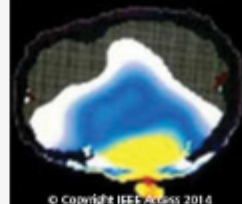
The Journal for rapid open access publishing

Would you give your child a cell phone?

**Comment Now**

on this controversial topic in IEEE Access.

5 Yr Old



© copyright IEEE Access 2014

**Comment**

The journal for rapid open access publishing.





# It's impossible to conduct research without software, say 7 out of 10 UK researchers

By [Simon Hettrick](#), Deputy Director.

No one knows how much software is used in research. Look around any lab and you'll see software – both standard and bespoke – being used by all disciplines and seniorities of researchers. Software is clearly fundamental to research, but we can't prove this without evidence. And this lack of evidence is the reason why we ran a survey of researchers at 15 Russell Group universities to find out about their software use and background.



## Headline figures

- 92% of academics use research software
- 69% say that their research would not be practical without it
- 56% develop their own software (worryingly, 21% of those have no training in software development)
- 70% of male researchers develop their own software, and only 30% of female researchers do so

## Also from Simon Hettrick

- 1 [It's impossible to conduct research without software, say 7 out of 10 UK researchers](#)
- 2 [Researchers both rely on software - and overlook it](#)
- 3 [From benign dictatorship to democratic association: the RSE AGM](#)
- 4 [The 20-line script that saves you hours of mind-numbing tedium](#)
- 5 [Anyone tried to explain research software to a teenager?](#)

## Page Tags

- [Size of research software community](#)
- [Policy research](#)
- [Surveys](#)

## Most Popular

- 1 [Software Evaluation Guide](#) – By Mike Jackson, Steve Crouch and Rob Baxter How do...

# Software Quality and Software Sustainability

- Open source is not a panacea! Commercial open source projects can produce high quality software but too often scientific software is of poor quality, undocumented and unmaintained
- The NSF now recognizes the importance of developing high quality maintainable scientific software in its SI<sup>2</sup> program:
  - **1. Scientific Software Elements (SSE):** SSE awards target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.
  - **2. Scientific Software Integration (SSI):** SSI awards target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems faced by NSF researchers in one or more areas of science and engineering. SSI awards will result in a sustainable community software framework serving a diverse community or communities.
  - **3. Scientific Software Innovation Institutes (S2I2):** S2I2 awards will focus on the establishment of long-term hubs of excellence in software infrastructure and technologies, which will serve a research community of substantial size and disciplinary breadth.

# TEACHING LAB SKILLS FOR SCIENTIFIC COMPUTING



## Who We Are

The [Software Carpentry Foundation](#) is a non-profit volunteer organization whose [members](#) teach researchers basic software skills.

*Learn more about [our history](#) and [our supporters](#).*

## What We Do

We run [over a hundred workshops a year](#), build and maintain [open access teaching materials](#), and run an [instructor training program](#).

*Find a [workshop](#) or [explore our lessons](#).*

## Get Involved

You can [host a workshop](#), [become an instructor](#), [support us](#), [help improve our lessons](#), [join our members' projects](#), or [join the discussion](#).

*Meet our [instructors](#) or [get in touch](#).*

## April 21 - 27, 2015: The People Behind Software Carpentry, Debating Scientific Software, Learning Objects, and Ally Skills Workshops.

By Anelda van der Walt / 2015-04-27

### Highlights

- Get to know the people behind Software Carpentry. First up: [Matt Davis](#).

### Conversations



## DATA CARPENTRY

MAKING DATA SCIENCE MORE EFFICIENT

Our sibling organization [Data Carpentry](#) teaches basic concepts, skills, and tools for working more effectively with data.

# **Data Science in the Future?**



254,000 RESULTS

[The \*\*Data Scientist\*\* role is a role of the future!](#)

[www.datascientists.net](http://www.datascientists.net) ▾

The **Data Scientist** role is a role of the future! Future proof your career and start transitioning today.

[Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...](#)

[jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job...](http://jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job...) ▾

Aug 10, 2011 · **Data scientists** are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

[LinkedIn's Monica Rogati On "What Is A \*\*Data Scientist\*\*?" - Forbes](#)

[www.forbes.com/.../linkedins-monica-rogati-on-what-is-a-data-scientist](http://www.forbes.com/.../linkedins-monica-rogati-on-what-is-a-data-scientist) ▾

Nov 27, 2011 · To continue our series on the emerging role of the **data scientist** in today's data-driven organizations, we spoke with Monica Rogati, Senior Data ...

Related searches for "**data scientist**"

[Data Scientist Seattle](#)

[Data Scientist Fortune](#)

[Data Scientist Salary](#)

[Data Scientist Jobs](#)

[Data Scientist Interview Ques...](#)

[Introduction to Data Science](#)

[Data scientist: The hot new gig in tech - Fortune Tech](#)

[tech.fortune.cnn.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech](http://tech.fortune.cnn.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech) ▾

Sep 06, 2011 · Companies that want to make sense of all their bits and bytes are hiring so-called **data scientists** - if they can find any. FORTUNE -- The unemployment rate ...

[The \*\*Data Scientist\*\* | Mine, Visualize, and Learn](#)

[www.thedatascientist.com](http://www.thedatascientist.com) ▾

As I jumped from room to room on Turntable.fm last night my eyes caught a glimpse of a rare room titled "AOKIxSOLREPUBLIC". I clicked it with a fury.

# 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 a 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.

**Conclusions?**

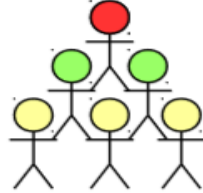
# Some final thoughts

- Computational research reproducibility is complicated!
- Open science needs access to software and data but are executable papers the answer?
- Making all research data open is not sensible or feasible e.g. LHC experiments
- Maintaining persistent links between publications and data is not easy
- Certainly need better training for research software developers and data scientists
- Need for culture change in university research ecosystem to recognize and provide career paths for research scientists who specialize in software development and data science



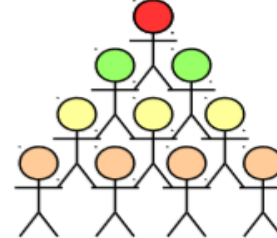
# How do we work?

## 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.

See opinion piece in Nature Physics last month ...

<http://www.nature.com/nphys/journal/v11/n5/full/nphys3313.html>

Available open access!

**Thank you for listening!**