The Fourth Paradigm: Data-Intensive Scientific Discovery and Open Science

Tony Hey

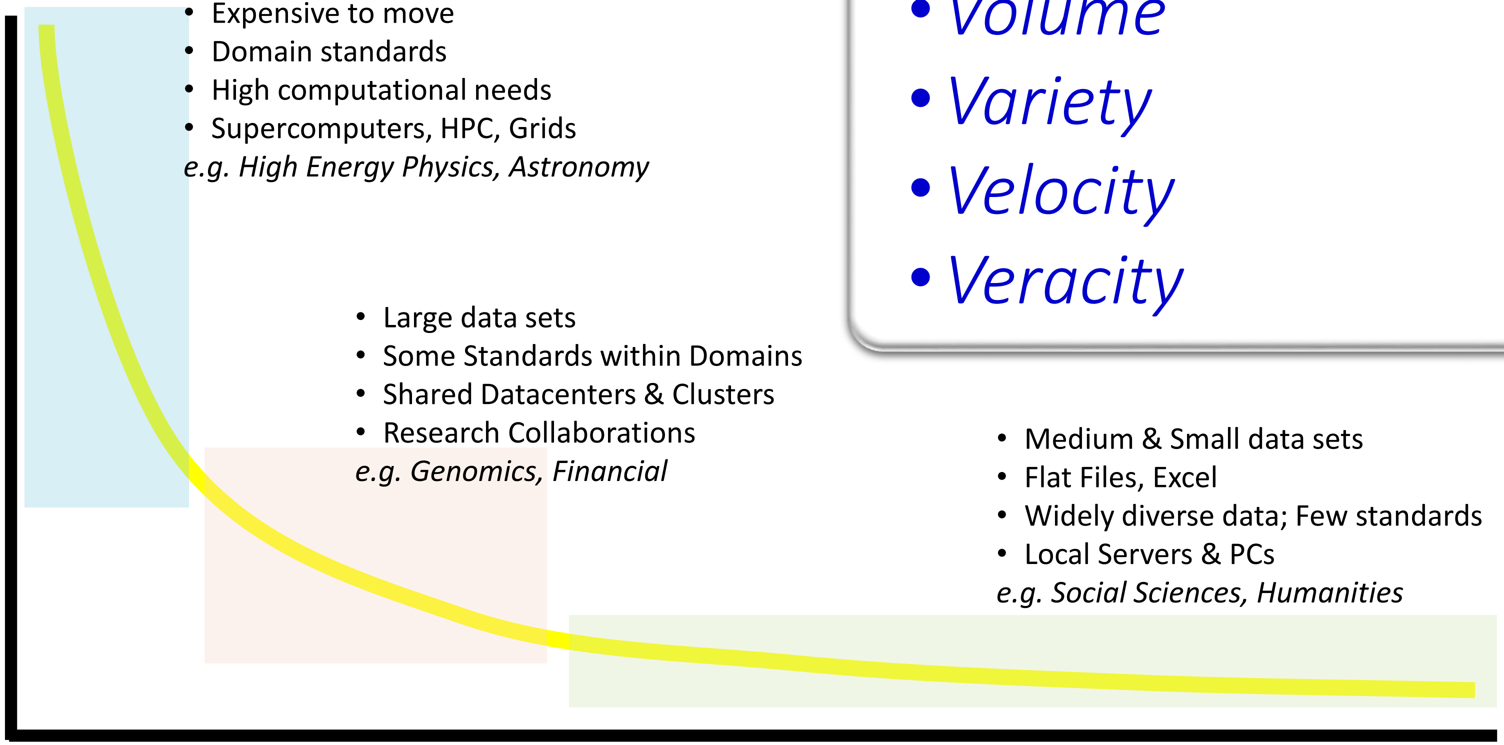
Chief Data Scientist

UK Science and Technology Facilities Council

tony.hey@stfc.ac.uk

The Fourth Paradigm: Data-Intensive Science

Much of Science is now Data-Intensive



- Extremely large data sets
- Expensive to move
- Domain standards
- High computational needs
- Supercomputers, HPC, Grids

e.g. High Energy Physics, Astronomy

- Large data sets
- Some Standards within Domains
- Shared Datacenters & Clusters
- Research Collaborations

e.g. Genomics, Financial

- Medium & Small data sets
- Flat Files, Excel
- Widely diverse data; Few standards
- Local Servers & PCs

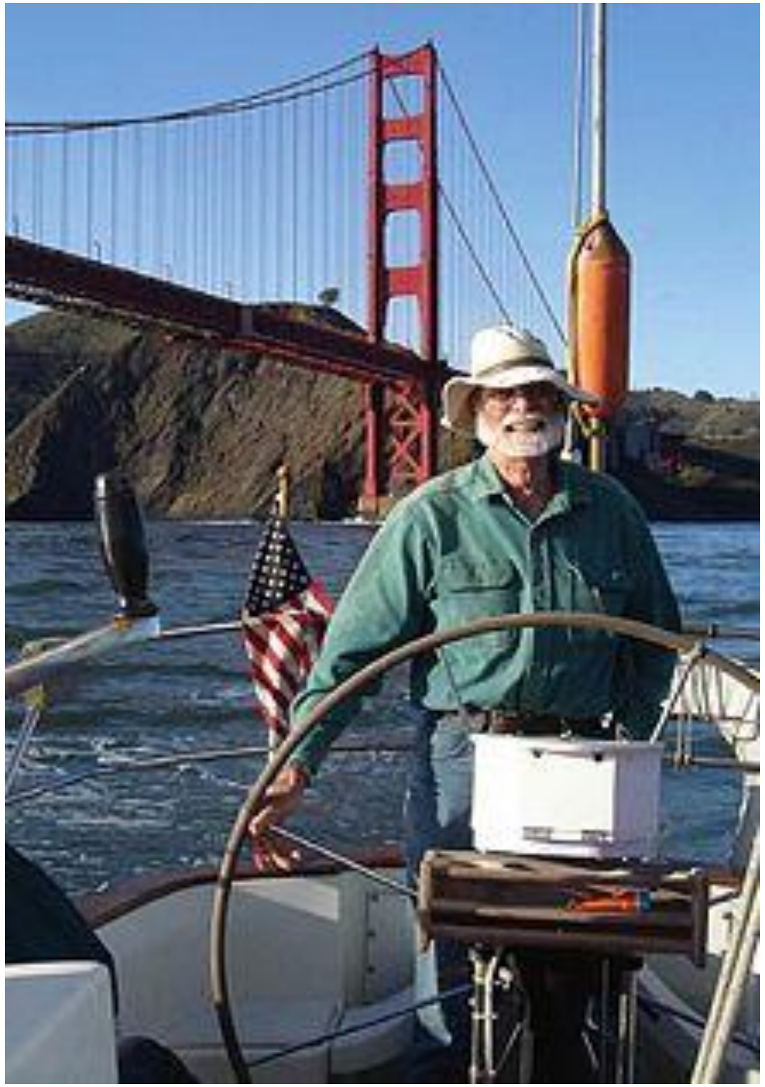
e.g. Social Sciences, Humanities

‘The Long Tail of Science’

Four “V’s” of Data

- *Volume*
- *Variety*
- *Velocity*
- *Veracity*

Jim Gray, Turing Award Winner



The ‘Cosmic Genome Project’: The Sloan Digital Sky Survey

- Survey of more than ¼ of the night sky
- Survey produces 200 GB of data per night
- Two surveys in one – images and spectra
- Nearly 2M astronomical objects, including 800,000 galaxies, 100,000 quasars
- 100’s of TB of data, and data is public
- Started in 1992, ‘finished’ in 2008

➢ The SkyServer Web Service was 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

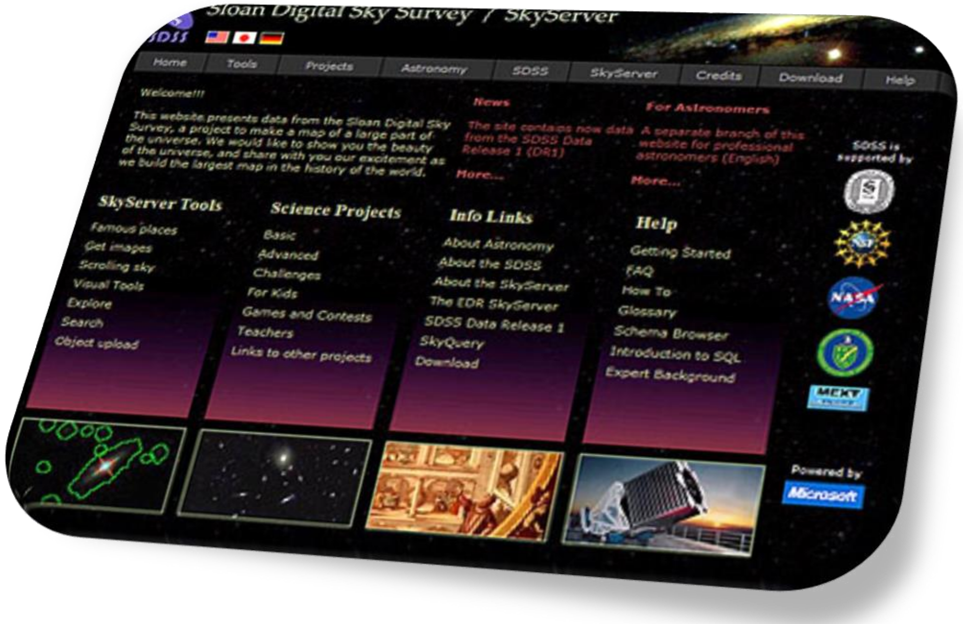
Open Data: Public Use of the Sloan Data

Posterchild in 21st century data publishing

- SkyServer web service has had over 400 million web
- About 1M distinct users vs 10,000 astronomers
- >1600 refereed papers!
- Delivered 50,000 hours of lectures to high schools

➢ New publishing paradigm: data is published <u>before</u> analysis by astronomers

➢ Platform for ‘citizen science’ with GalaxyZoo project



Welcome!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site contains now data from the SDSS Data Release 1 (DR1)

More...

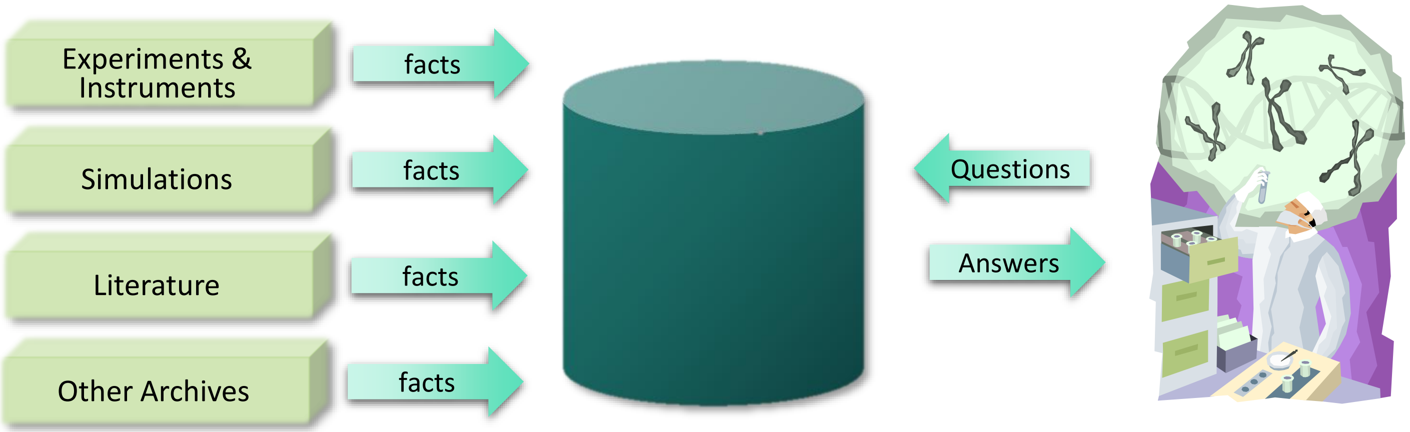
For Astronomers

A separate branch of this website for professional astronomers (English)

More...

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



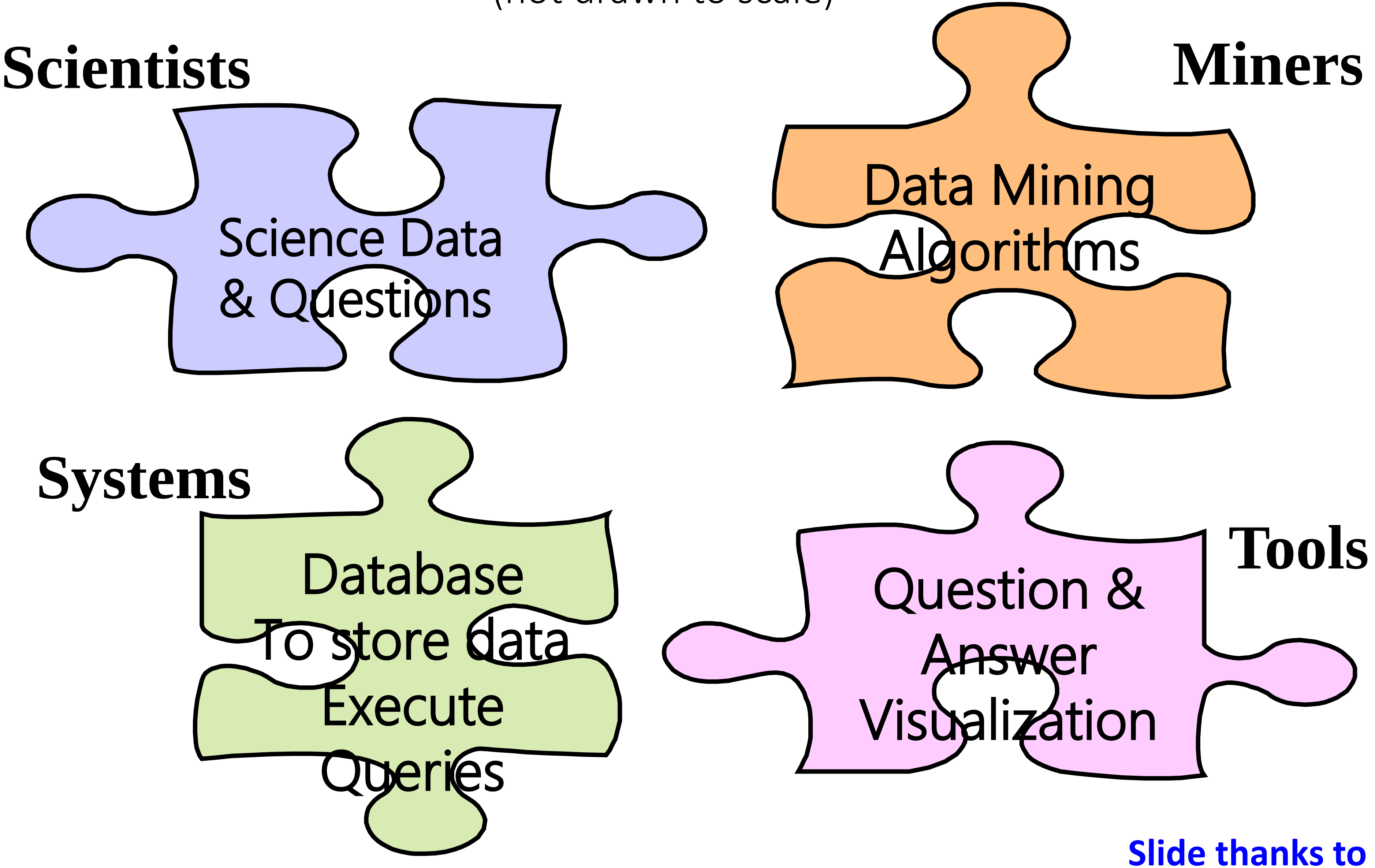
The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

Slide thanks to Jim Gray

What X-info Needs from Computer Science

(not drawn to scale)



Scientists

Science Data & Questions

Miners

Data Mining Algorithms

Systems

Database
To store data
Execute
Queries

Tools

Question & Answer Visualization

Slide thanks to Jim Gray

Working Cross-Culture: A Way to Engage With Domain Scientists

- Communicate in terms of scenarios
- Work on a problem that gives 100x benefit
    - Weeks/task vs hours/task
- Solve 20% of the problem
    - The other 80% will take decades
- Prototype
- Go from working-to-working: Always have
    - Something to show
    - Clear next steps
    - Clear goal
- Avoid death-by-collaboration-meetings

Slide thanks to Jim Gray

The Fourth Paradigm: Data-Intensive Science

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

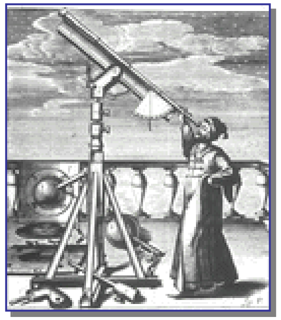
- Newton's Laws, Maxwell's Equations…

Last few decades – **Computational Science**

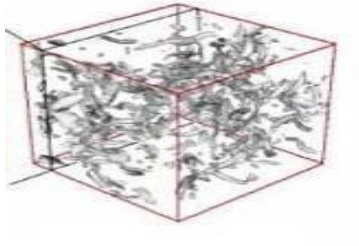
- Simulation of complex phenomena

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



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

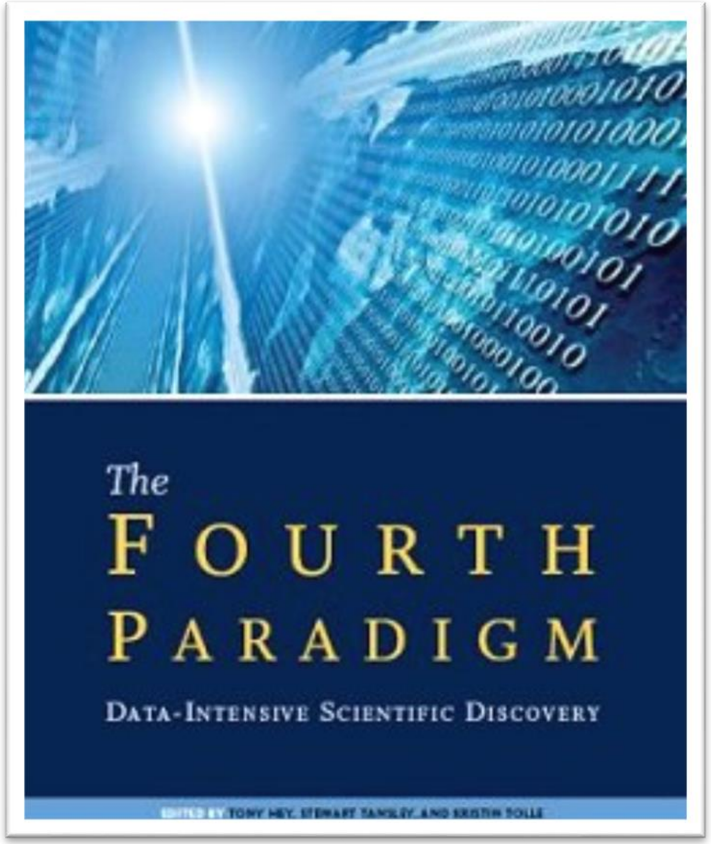


eScience is the set of tools and technologies to support data federation and collaboration

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

With thanks to Jim Gray

Data-Intensive Scientific Discovery


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

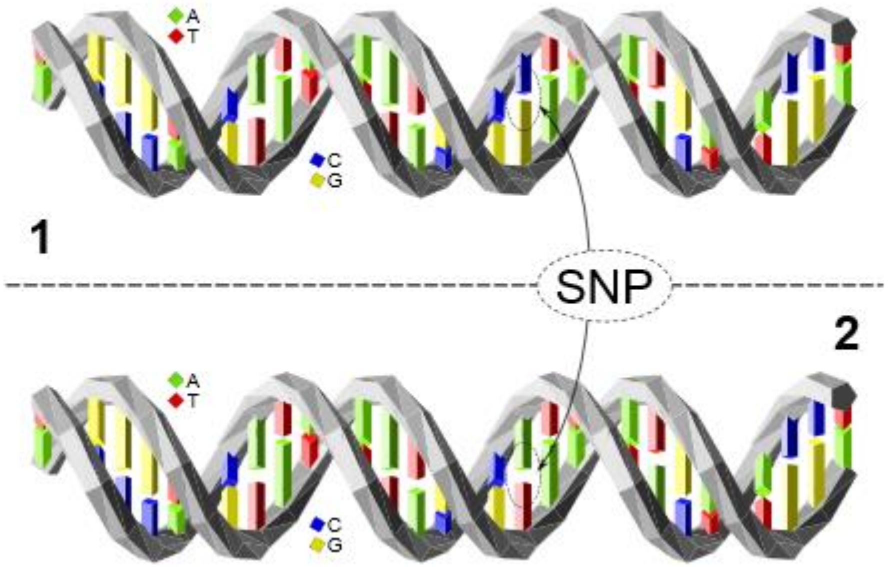
Published under Creative Commons License and available online from The Fourth Paradigm and Science@Microsoft at http://research.microsoft.com and on Amazon.com

Three Examples of Data-Intensive Science

Genomics and Personalized medicine

Use genetic markers (e.g. SNPs) to…

- ➢ Understand causes of disease
- ➢ Diagnose a disease
- ➢ Infer propensity to get a disease
- ➢ Predict reaction to a drug

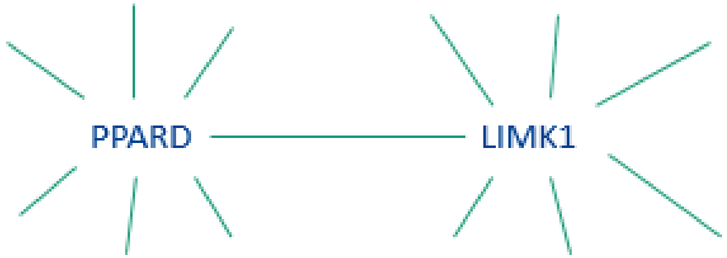


Genomics, Machine Learning and the Cloud

The Problem

- Wellcome Trust data for seven common diseases
- Look at all SNP pairs (about 60 billion)
- Analysis with state-of-the-art Machine Learning algorithm requires 1,000 compute years and produces 20 TB output
- Using 27,000 compute cores in Microsoft's Cloud, the analysis was completed in 13 days

First result: SNP pair implicated in coronary artery disease

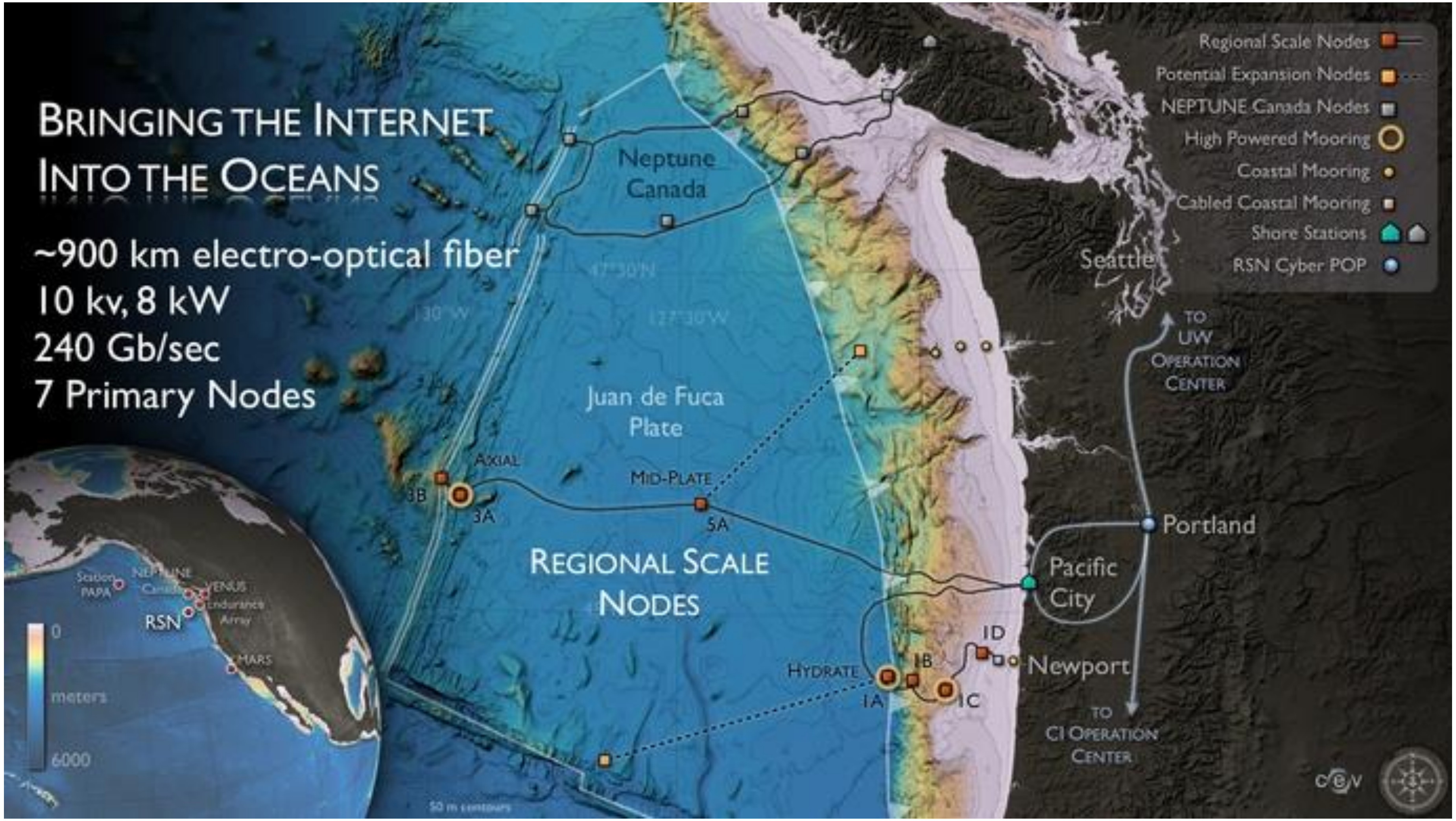


SCIENTIFIC REPORTS

An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data

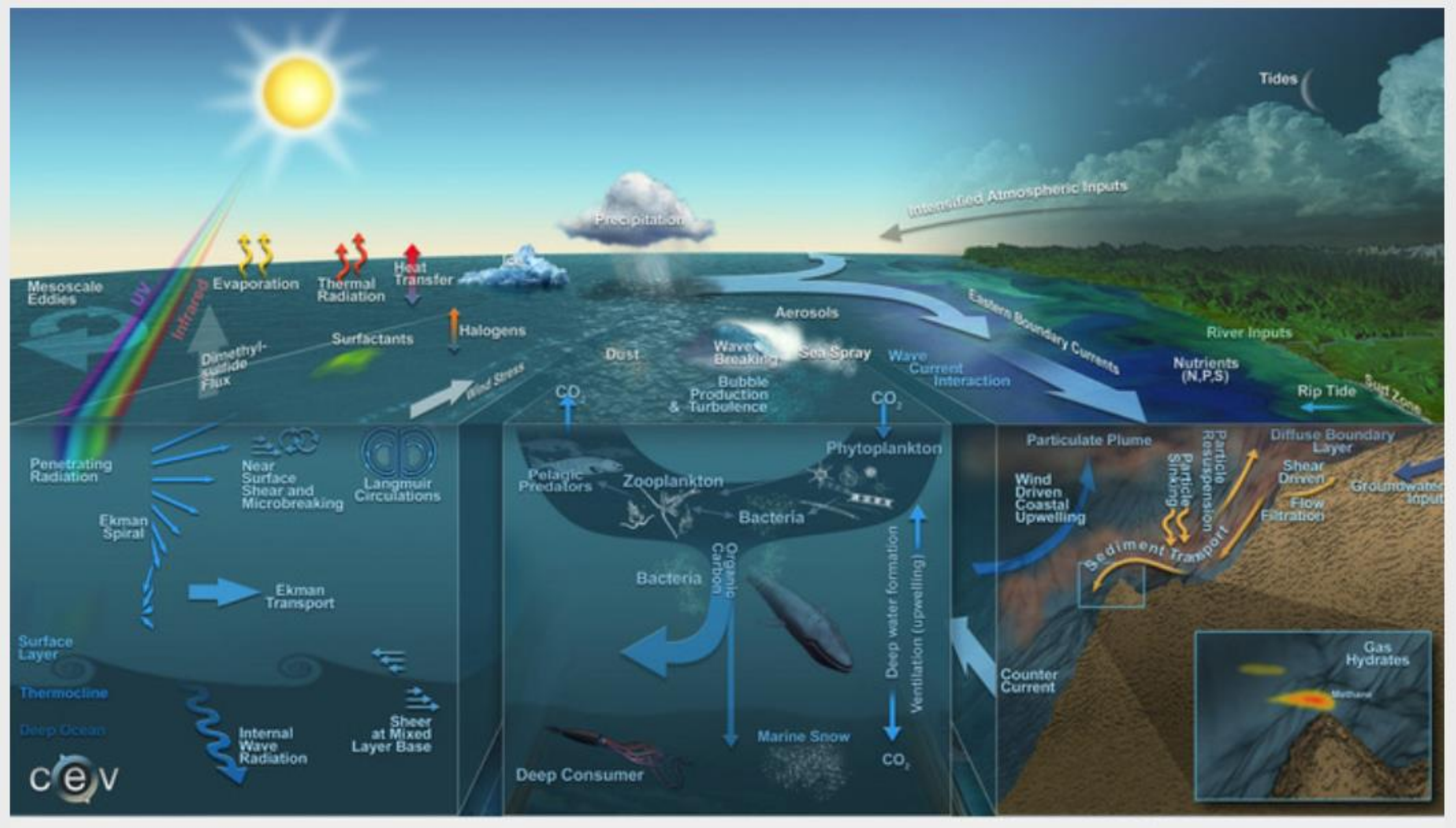
Christoph Lippert, Jennifer Listgarten, Robert I. Davidson, Jeff Baxter, Hoifung Poon, Carl M. Kadie & David Heckerman

NSF’s Ocean Observatory Initiative



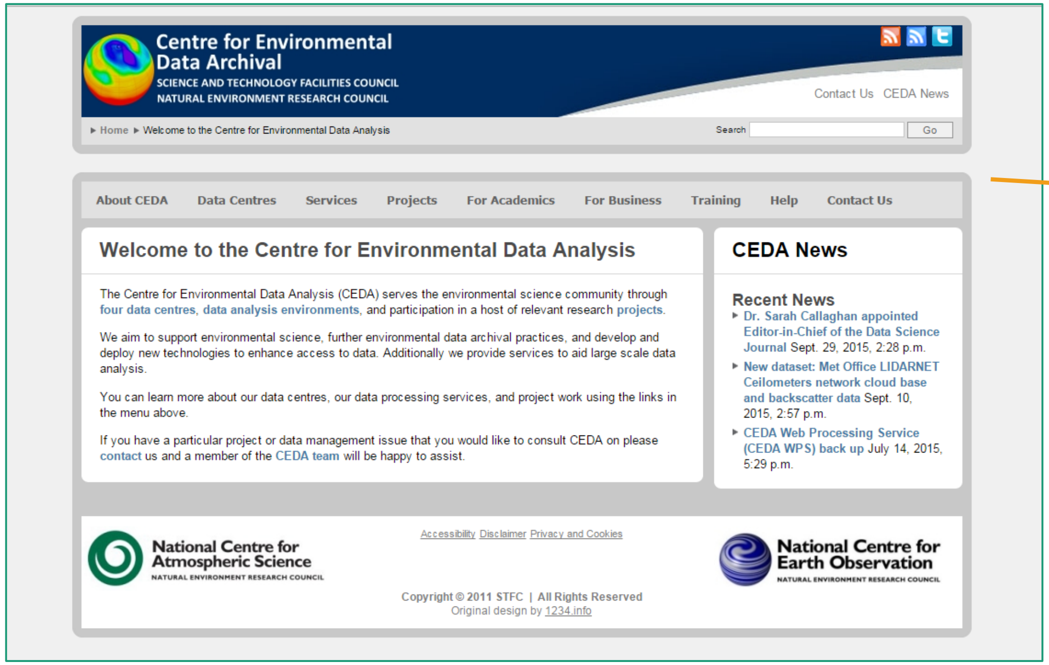
Slide courtesy of John Delaney

Oceans and Life



Slide courtesy of John Delaney

CEDA: Centre for Environmental Data Analysis


Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL
Contact Us
CEDA News
▸ Home ▸ Welcome to the Centre for Environmental Data Analysis
Search
Go
About CEDA
Data Centres
Services
Projects
For Academics
For Business
Training
Help
Contact Us

Welcome to the Centre for Environmental Data Analysis

The Centre for Environmental Data Analysis (CEDA) serves the environmental science community through four data centres, data analysis environments, and participation in a host of relevant research projects.

We aim to support environmental science, further environmental data archival practices, and develop and deploy new technologies to enhance access to data. Additionally we provide services to aid large scale data analysis.

You can learn more about our data centres, our data processing services, and project work using the links in the menu above.

If you have a particular project or data management issue that you would like to consult CEDA on please contact us and a member of the CEDA team will be happy to assist.

CEDA News

Recent News

- Dr. Sarah Callaghan appointed Editor-in-Chief of the Data Science Journal Sept. 29, 2015, 2:28 p.m.
- New dataset: Met Office LIDARNET Ceilometers network cloud base and backscatter data Sept. 10, 2015, 2:57 p.m.
- CEDA Web Processing Service (CEDA WPS) back up July 14, 2015, 5:29 p.m.

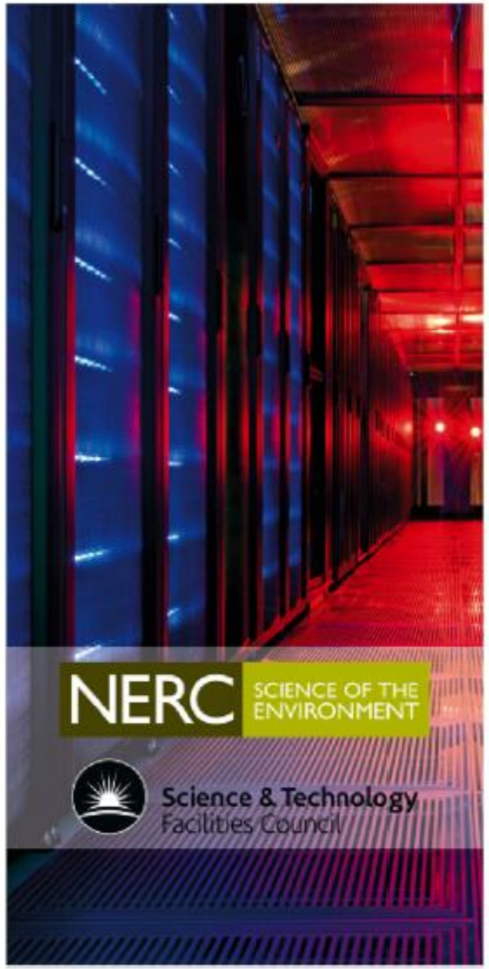
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL
Accessibility Disclaimer Privacy and Cookies
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Copyright © 2011 STFC | All Rights Reserved
Original design by 1234.info

To support environmental science, further environmental data archival practices, and develop and deploy new technologies to enhance access to data

Centre for Environmental Data Analysis: JASMIN infrastructure

Part data store, part supercomputer, part private cloud…


NERC SCIENCE OF THE ENVIRONMENT
Science & Technology Facilities Council

- 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.

End-to end Network Support for Data-intensive Research?

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:

7. Supercomputing and HPC resources
8. Training of Scientific Software Engineers and Data Scientists

NSF Task Force on ‘Campus Bridging’ (2011)

The goal of ‘campus bridging’ is to enable the seamlessly integrated use among:

- a researcher’s personal cyberinfrastructure
- cyberinfrastructure at other campuses
- cyberinfrastructure at the regional, national and international levels

so that they all function as if they were proximate to the scientist

National Science Foundation
Advisory Committee for CyberInfrastructure
Task Force on Campus Bridging

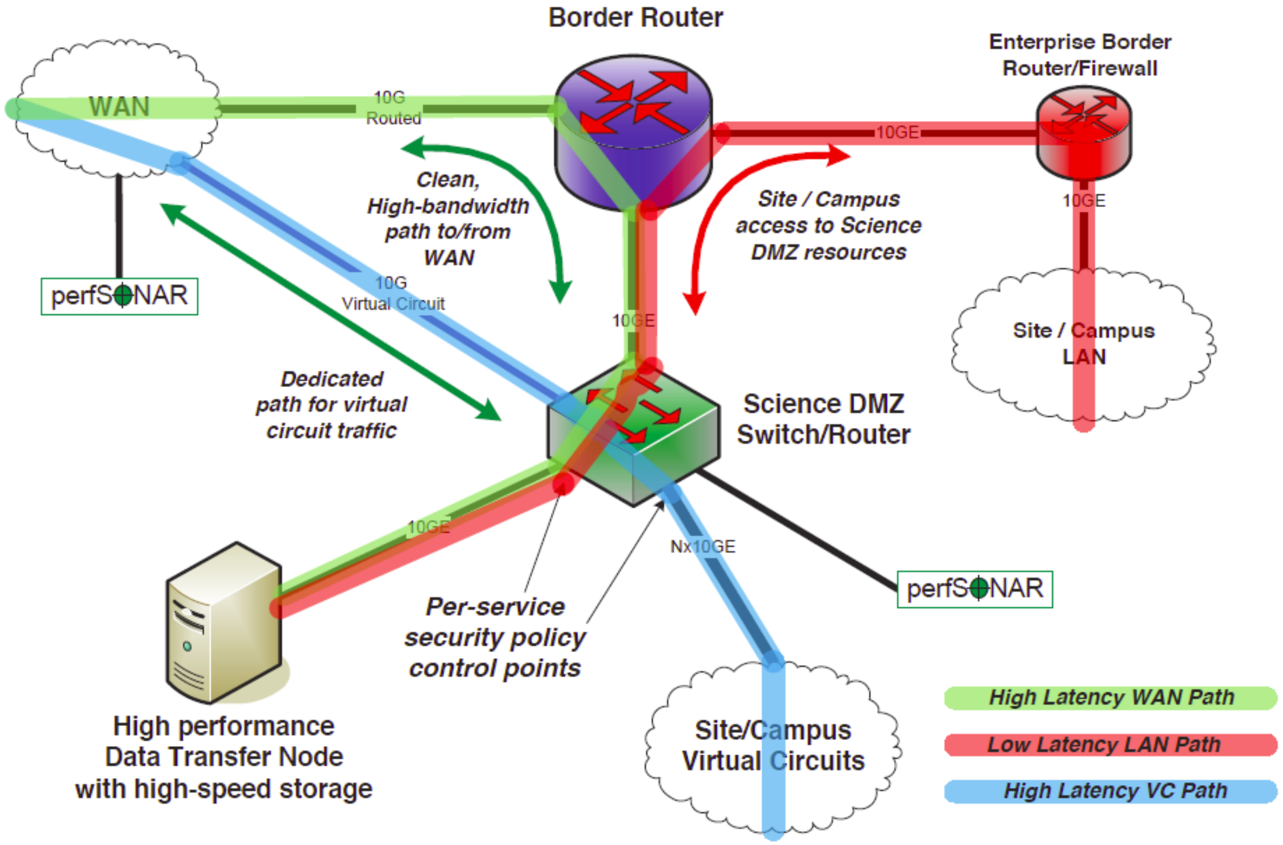
Final Report, March 2011

What are ‘Science DMZs’ and why do we need them?

- The Science DMZ model addresses network performance problems seen at research institutions

- It creates an environment optimized for data-intensive scientific applications such as high volume bulk data transfer or remote control of experiments

- Most networks designed to support general-purpose business operations and are not capable of supporting the data movement requirements of data-intensive science applications

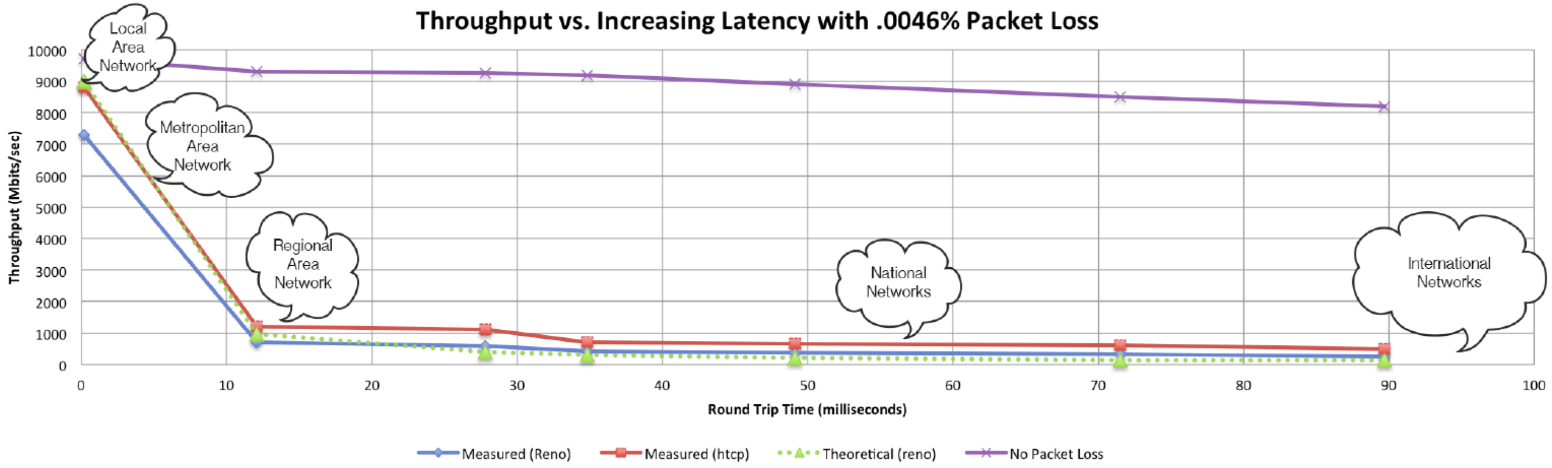


Border Router
Enterprise Border Router/Firewall
WAN
10G Routed
10GE
Clean, High-bandwidth path to/from WAN
Site / Campus access to Science DMZ resources
10GE
perfSONAR
10G Virtual Circuit
10GE
Site / Campus LAN
Dedicated path for virtual circuit traffic
Science DMZ Switch/Router
10GE
Nx10GE
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage
Site/Campus Virtual Circuits
High Latency WAN Path
Low Latency LAN Path
High Latency VC Path

Thanks to Eli Dart, LBNL

The Problem of Packet Loss

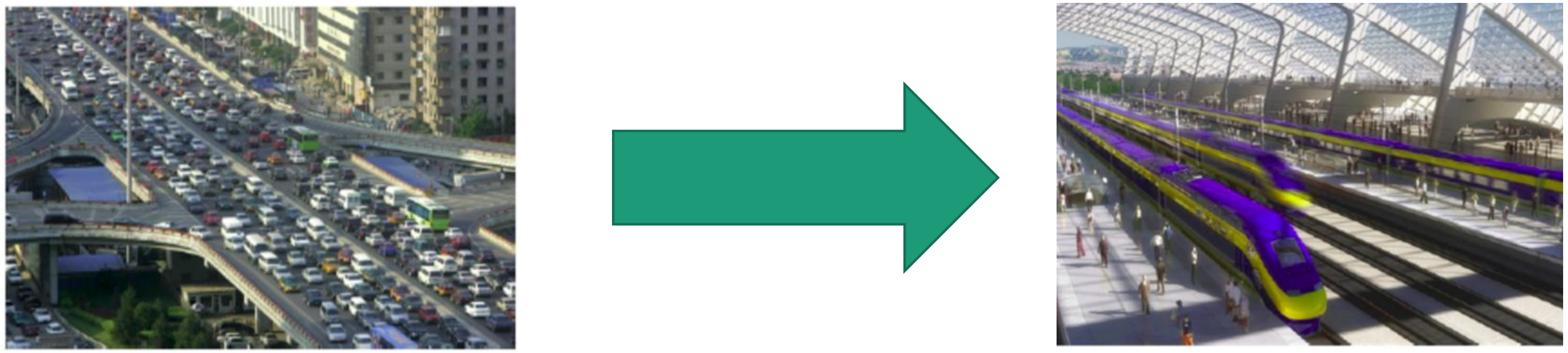
- Most scientific data transfers use TCP
- Packet loss can cause dramatic loss in throughput
- TCP interprets packet loss as network congestion and reduces rate of transmission of data



- ➢The Science DMZ model provides the framework for building a network infrastructure that is more loss tolerant

Thanks to Eli Dart, LBNL

Need for European adoption of ‘Science DMZ’ end-to-end network architecture



- Science DMZs implemented at over 100 US universities
- NSF invested more than $60M in DMZ campus cyberinfrastructure

➢ Need to connect ESFRI Large Experimental Facilities and HPC systems via Science DMZs

➢ Need research funding agencies to work together with GEANT and NRENs to support high bandwidth end-to-end connections to researchers at institutions

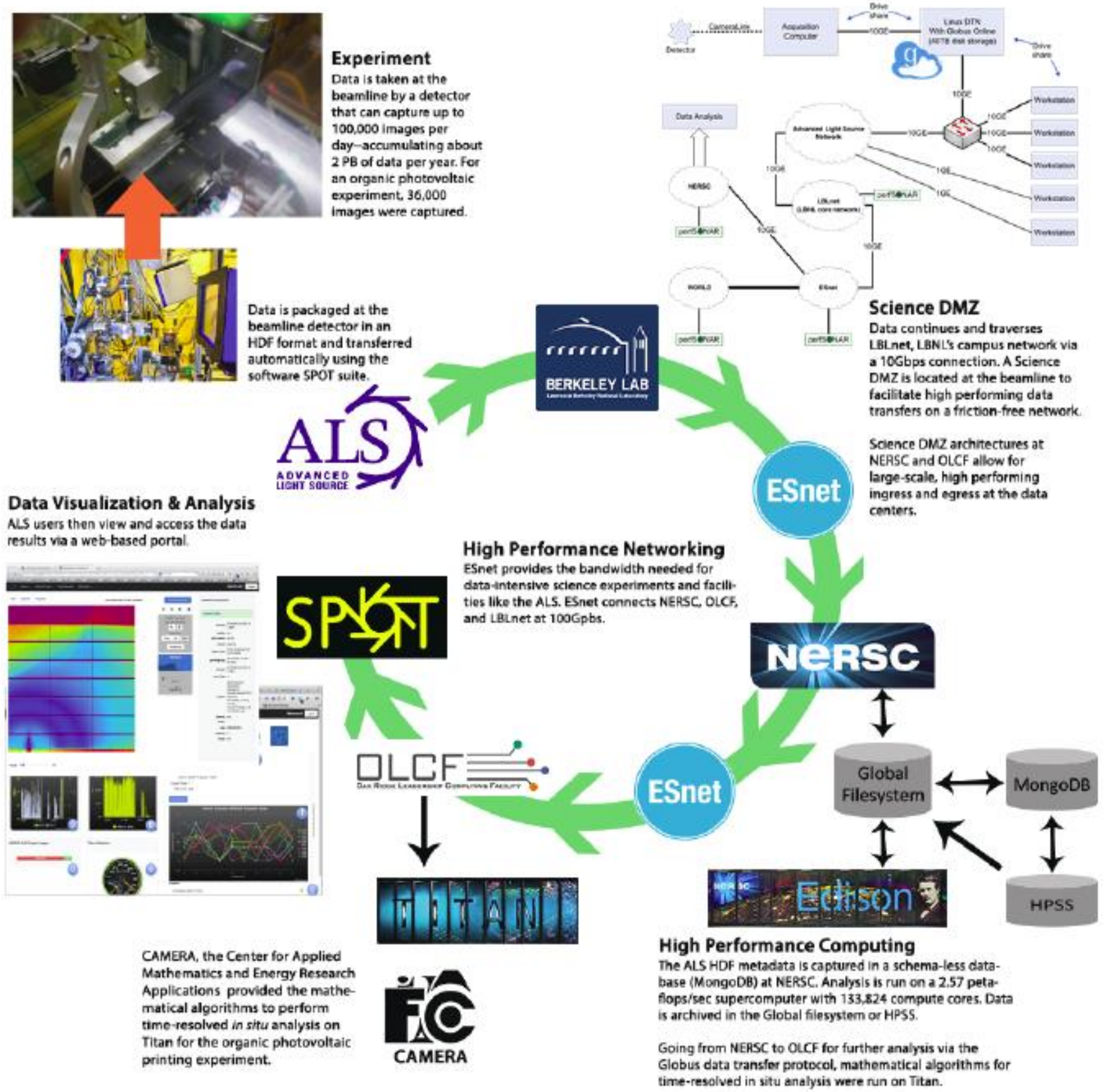
➢ AAI systems can support industry access to research infrastructure

Creation of European 'Superfacilities'?

- In the US large experimental facilities are creating 'superfacilities' to solve advanced science questions by tightly coupling distributed resources

- Data volume and analysis needs for many experiments are growing faster than the experimental facility computing resources

- Experimental facilities with the greatest data growth are integrating:
  - Remote HPC resources
  - Advanced workflow and analysis tools
  - High-performance networks capable of supporting data-intensive science



Experiment
Data is taken at the beamline by a detector that can capture up to 100,000 images per day—accumulating about 2 PB of data per year. For an organic photovoltaic experiment, 36,000 images were captured.

Data is packaged at the beamline detector in an HDF format and transferred automatically using the software SPOT suite.

Science DMZ
Data continues and traverses LBLnet, LBNL's campus network via a 10Gbps connection. A Science DMZ is located at the beamline to facilitate high performing data transfers on a friction-free network.

Science DMZ architectures at NERSC and OLCF allow for large-scale, high performing ingress and egress at the data centers.

Data Visualization & Analysis
ALS users then view and access the data results via a web-based portal.

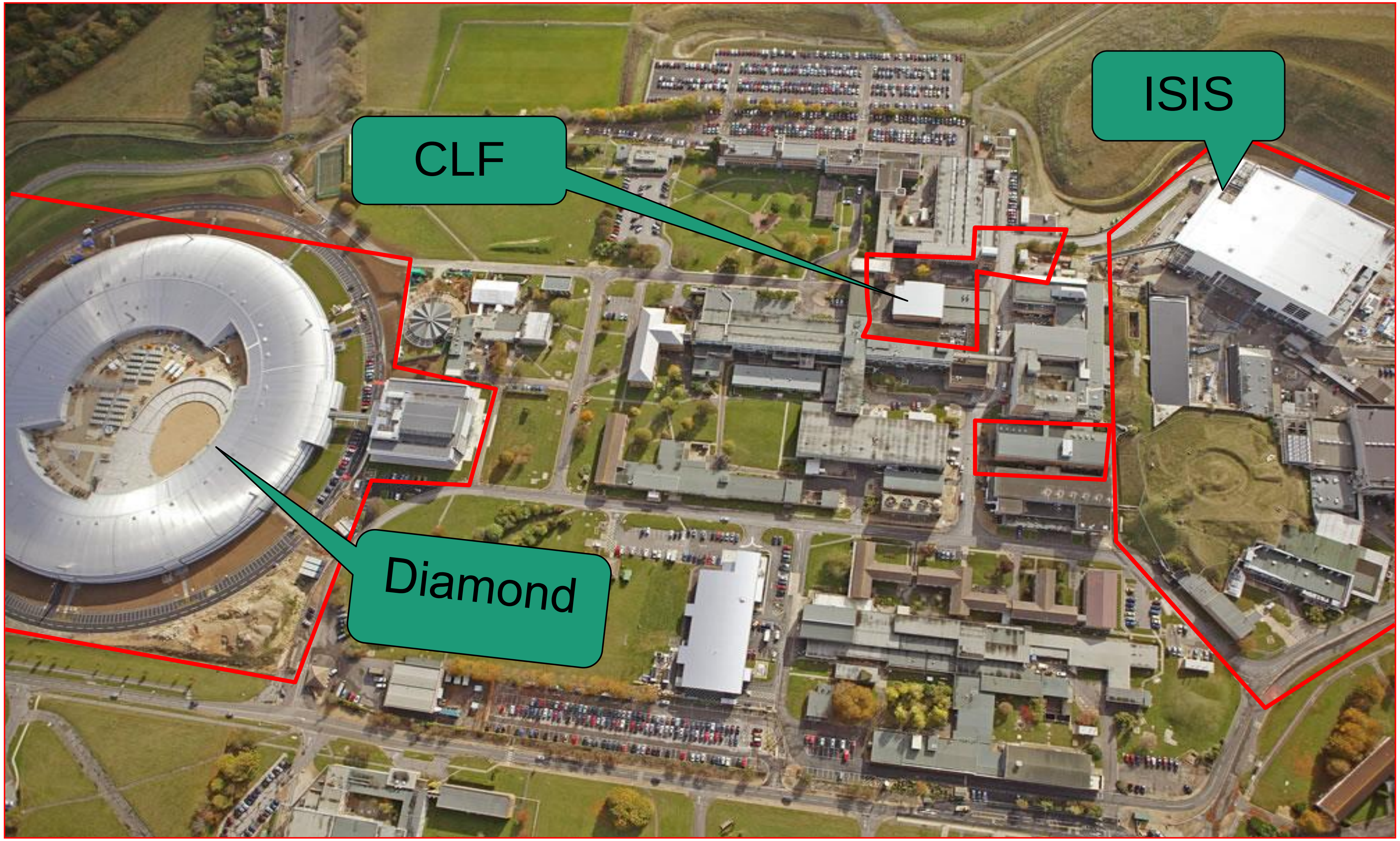
High Performance Networking
ESnet provides the bandwidth needed for data-intensive science experiments and facilities like the ALS. ESnet connects NERSC, OLCF, and LBLnet at 100Gpbs.

CAMERA, the Center for Applied Mathematics and Energy Research Applications provided the mathematical algorithms to perform time-resolved *in situ* analysis on Titan for the organic photovoltaic printing experiment.

High Performance Computing
The ALS HDF metadata is captured in a schema-less database (MongoDB) at NERSC. Analysis is run on a 2.57 petaflops/sec supercomputer with 133,824 compute cores. Data is archived in the Global filesystem or HPSS.

Going from NERSC to OLCF for further analysis via the Globus data transfer protocol, mathematical algorithms for time-resolved in situ analysis were run on Titan.

STFC Harwell Site Experimental Facilities in UK

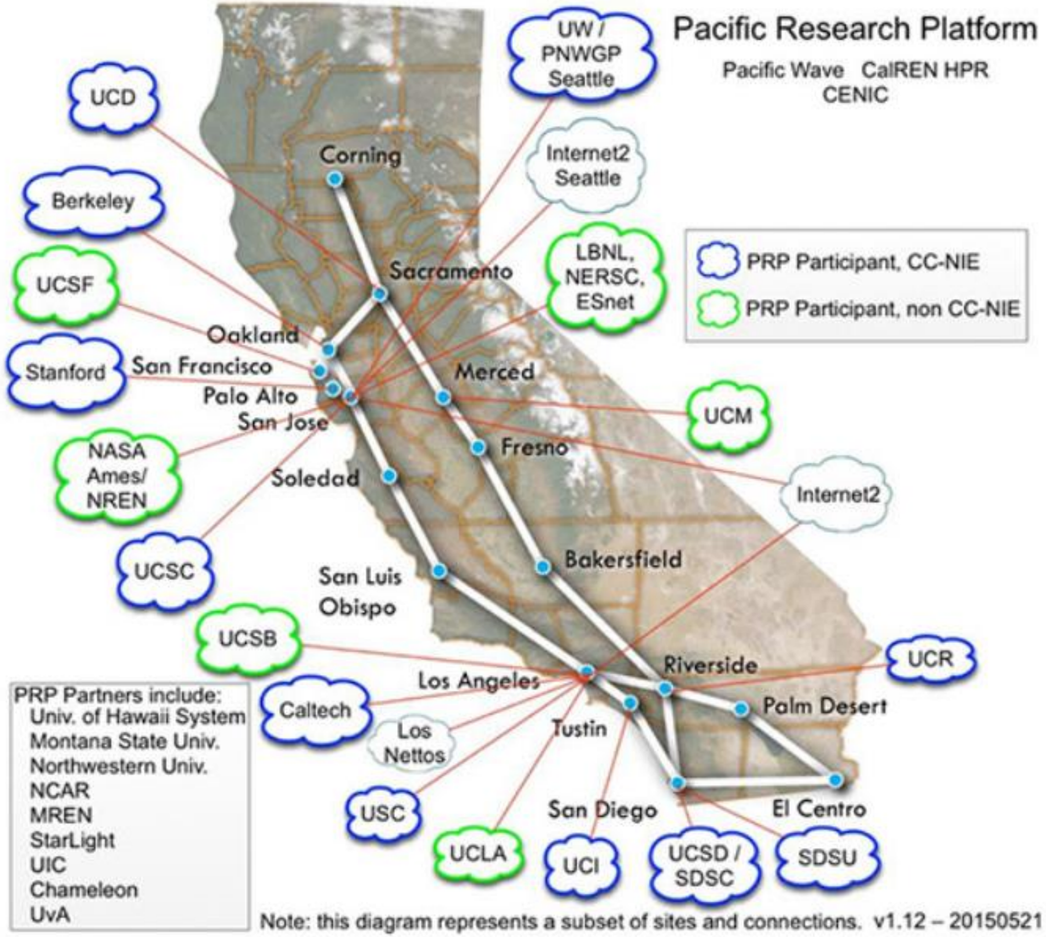


Pacific Research Platform

- NSF funding $5M award to UC San Diego and UC Berkeley to establish a science-driven high-capacity data-centric “freeway system” on a large regional scale.
- This network infrastructure will give the research institutions the ability to move data 1,000 times faster compared to speeds on today’s Internet.

August 2015

➢ “PRP will enable researchers to use standard tools to move data to and from their labs and their collaborators’ sites, supercomputer centers and data repositories distant from their campus IT infrastructure, at speeds comparable to accessing local disks,” said co-PI Tom DeFanti



The PRP partners are connected by CENIC’s 100G and 10G infrastructure as shown. CENIC is connected to DOE’s ESnet and Internet2 as well as Pacific Wave, all at 100G.

Open Science and the US OSTP Memo

US White House Memorandum on Increased Public Access to Research Results

- 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

Open Access: 2013 as the Tipping Point?

• US White House Memorandum	22 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	2 August 2013

University of California approves Open Access

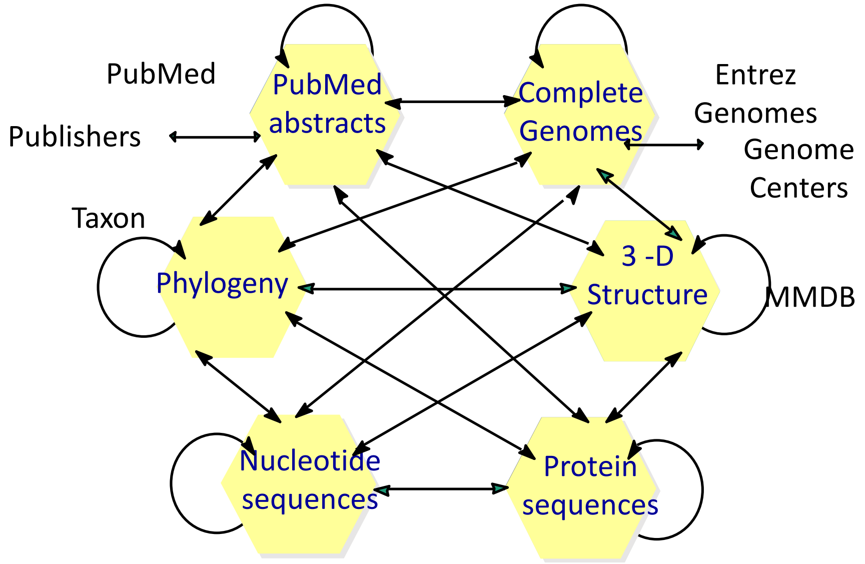
- UC is the largest public research university in the world and its faculty members receive roughly 8% of all research funding in the U.S.
- UC produces 40,000 publications per annum corresponding to about 2 – 3 % of all peer-reviewed articles in world each year
- UC policy requires all 8000 faculty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

2 August 2013

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.

PubMed Central



Entrez cross-database search

NIH Open Access Compliance?

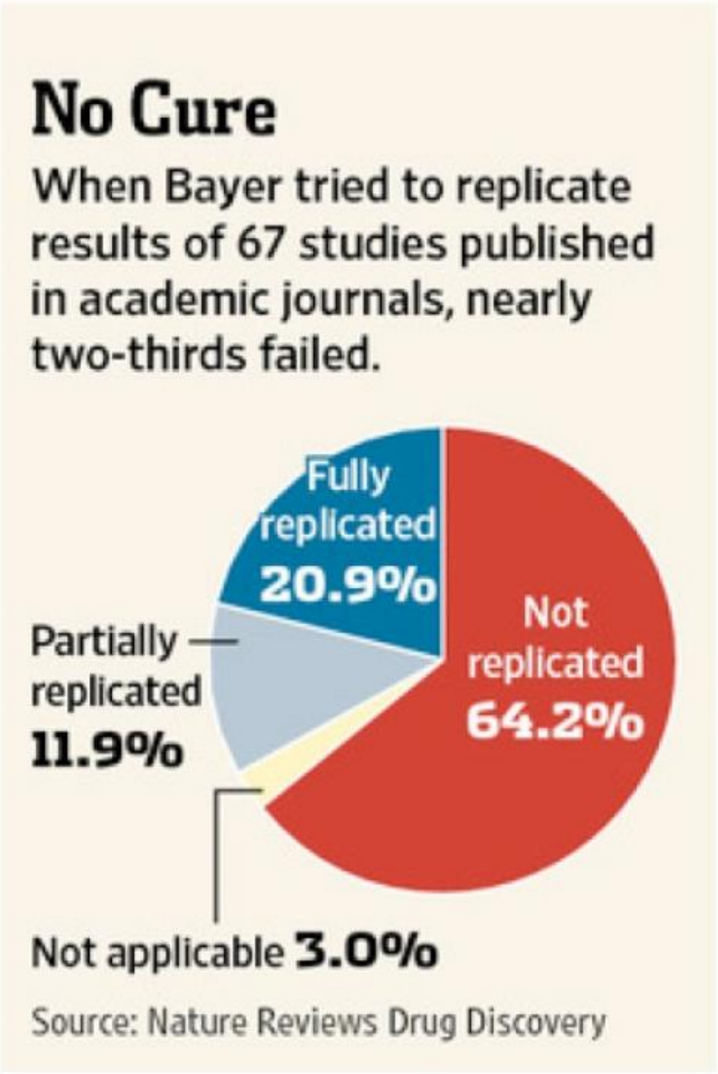
- PMC Compliance Rate
  - Before legal mandate compliance was 19%
  - Signed into law by George W. Bush in 2007
  - After legal mandate compliance up to 75%
- NIH have taken further step of announcing in 2013 that they

  '… will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.'
- Since NIH implemented their policy about continuation awards
  - Compliance rate increasing ½% per month
  - By November 2014, compliance rate had reached 86%

Serious problems of research reproducibility in bioinformatics

During a decade as head of global cancer research at Amgen, C. Glenn Begley identified 53 "landmark" publications -- papers in top journals, from reputable labs -- for his team to reproduce.

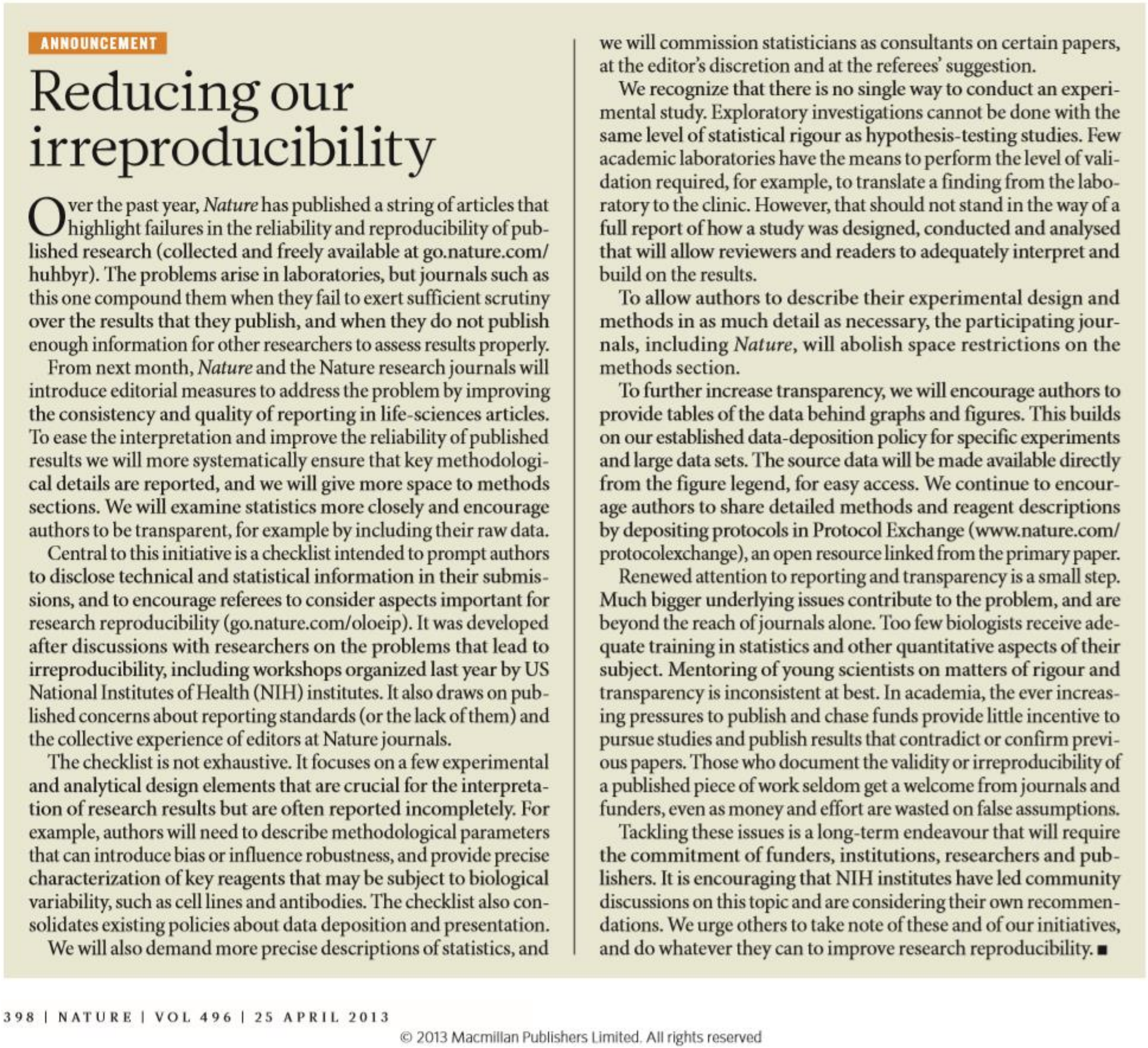
Result: 47 of the 53 could not be replicated!



No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discovery



ANNOUNCEMENT

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). 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/oloeip). 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.

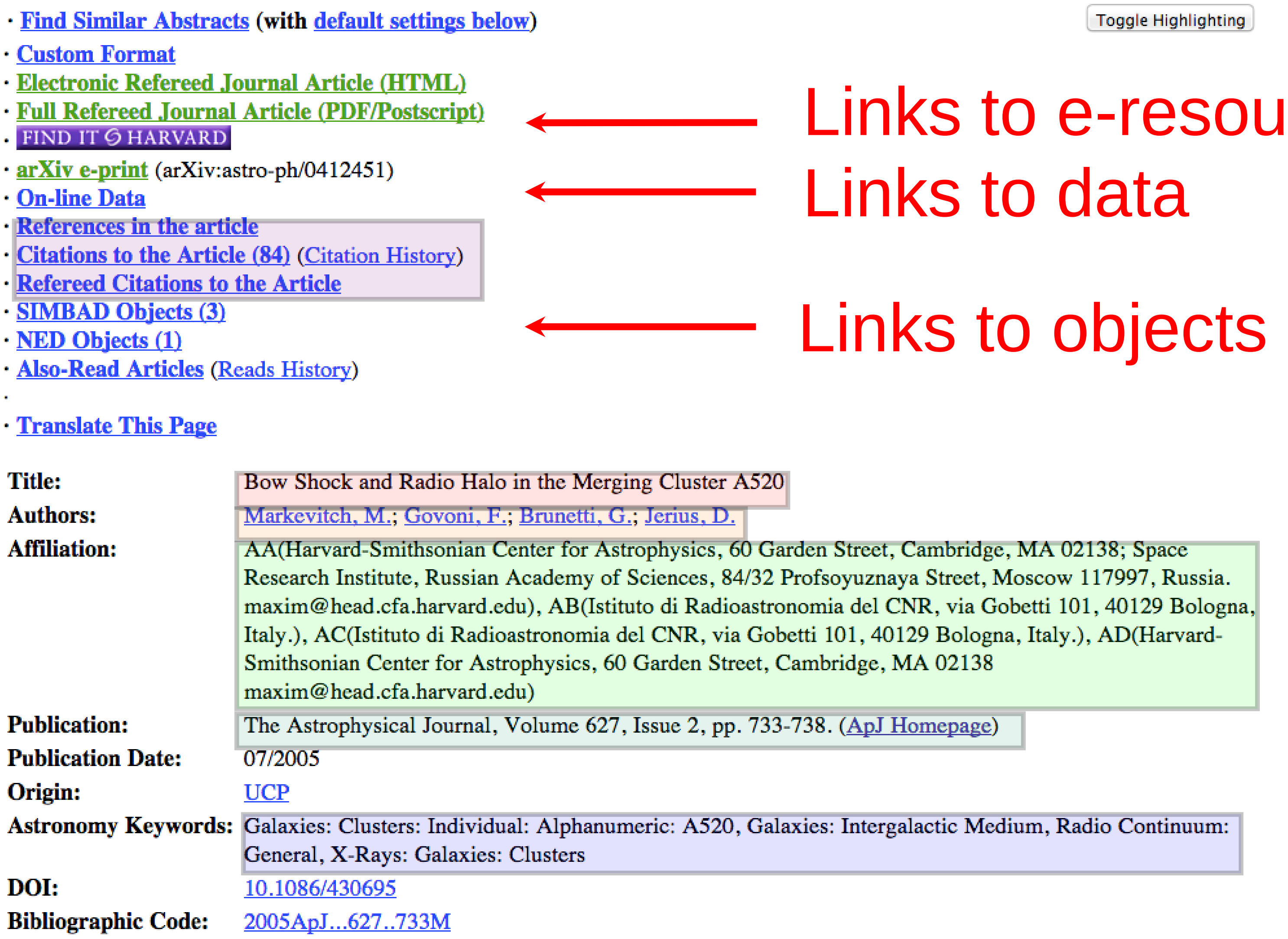
To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access.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

Linking Publications to Data: The State of the Art

Astrophysics Data System ADS



- Find Similar Abstracts (with default settings below)
- Custom Format
- Electronic Refereed Journal Article (HTML)
- Full Refereed Journal Article (PDF/Postscript)
- FIND IT @ HARVARD
- arXiv e-print (arXiv:astro-ph/0412451)
- On-line Data
- References in the article
- Citations to the Article (84) (Citation History)
- Refereed Citations to the Article
- SIMBAD Objects (3)
- NED Objects (1)
- Also-Read Articles (Reads History)
-
- Translate This Page

Title:	Bow Shock and Radio Halo in the Merging Cluster A520
Authors:	Markevitch, M.; Govoni, F.; Brunetti, G.; Jerius, D.
Affiliation:	AA(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; Space Research Institute, Russian Academy of Sciences, 84/32 Profsoyuznaya Street, Moscow 117997, Russia. maxim@head.cfa.harvard.edu), AB(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AC(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AD(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138 maxim@head.cfa.harvard.edu)
Publication:	The Astrophysical Journal, Volume 627, Issue 2, pp. 733-738. (ApJ Homepage)
Publication Date:	07/2005
Origin:	UCP
Astronomy Keywords:	Galaxies: Clusters: Individual: Alphanumeric: A520, Galaxies: Intergalactic Medium, Radio Continuum: General, X-Rays: Galaxies: Clusters
DOI:	10.1086/430695
Bibliographic Code:	2005ApJ...627..733M

Links to e-resources

Links to data

Links to objects

ourt

Literature

WIKIPEDIA
The Free Encyclopedia

AstroBetter

Blogs, Wikis, etc.

Courtesy of Alyssa Goodman

"Seamless Astronomy" (Tools)

ads labs NASA

ALADIN

World Wide Telescope

Astrometry.net

flickr

TOPCAT

ds9

Disclaimer: *This slide shows key excerpts from within the astronomy community & excludes more general s/w that is used, such as Papers, Zotero, Mendeley, EndNote, graphing & statistics packages, data handling software, search engines, etc.*

Data

VizieR

SIMBAD

AstroGrid

VAO

IVOA

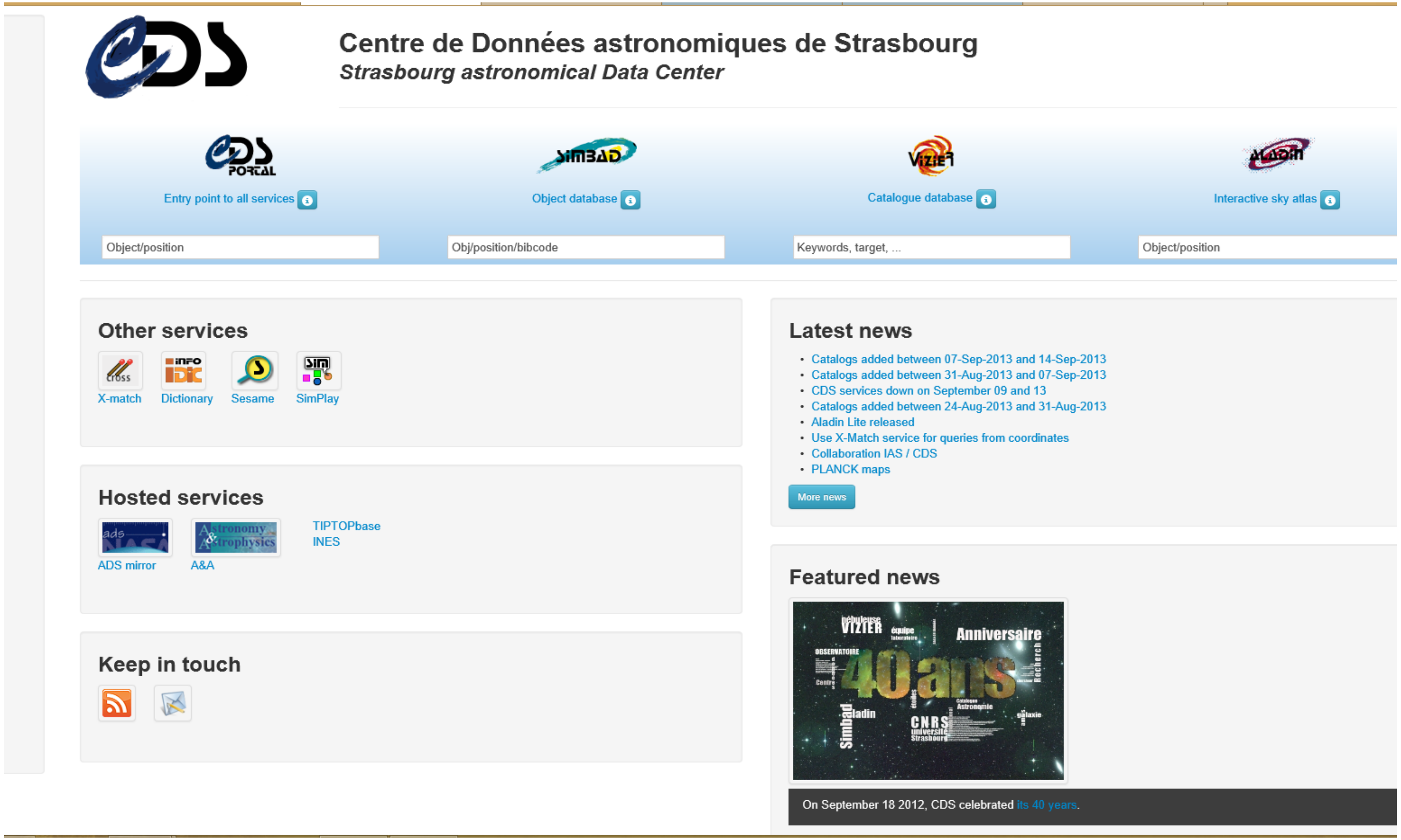
EURO VO

AIDA Astronomical Infrastructure for Data Access

"Registries"

DataScope

Strasbourg CDS Data Services



Centre de Données astronomiques de Strasbourg

Strasbourg astronomical Data Center

Entry point to all services

Object database

Catalogue database

Interactive sky atlas

Object/position

Obj/position/bibcode

Keywords, target, ...

Object/position

Other services

X-match

Dictionary

Sesame

SimPlay

Hosted services

ADS mirror

A&A

TIPTOPbase

INES

Keep in touch

Latest news

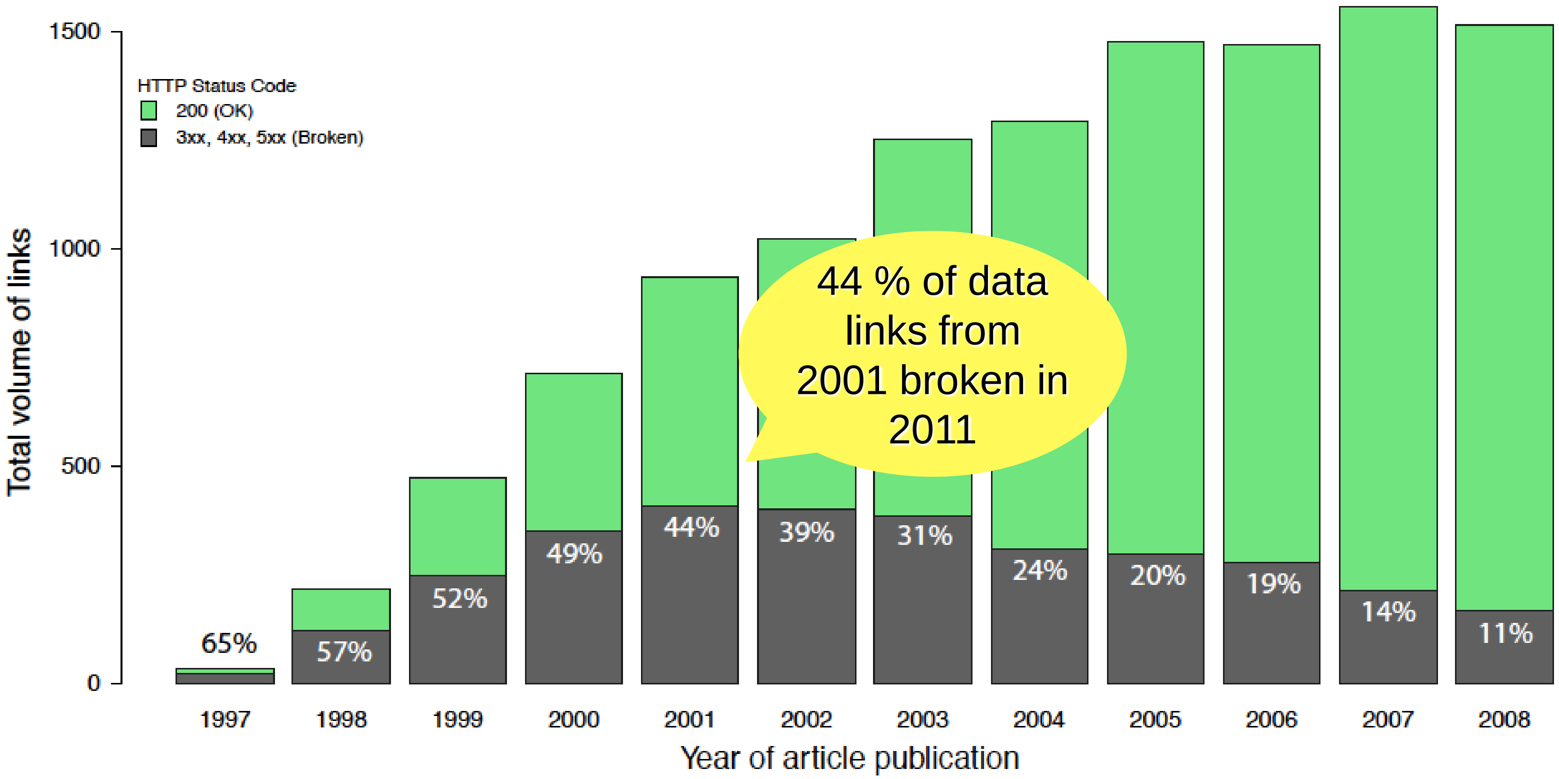
- Catalogs added between 07-Sep-2013 and 14-Sep-2013
- Catalogs added between 31-Aug-2013 and 07-Sep-2013
- CDS services down on September 09 and 13
- Catalogs added between 24-Aug-2013 and 31-Aug-2013
- Aladin Lite released
- Use X-Match service for queries from coordinates
- Collaboration IAS / CDS
- PLANCK maps

More news

Featured news

On September 18 2012, CDS celebrated its 40 years.

Sustainability of Data Links?



44 % of data links from 2001 broken in 2011

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. .

Pepe et al. 2012

Datacite and ORCID

DataCite

- International consortium to establish easier access to scientific research data
- Increase acceptance of research data as legitimate, citable contributions to the scientific record
- Support data archiving that will permit results to be verified and re-purposed for future study.

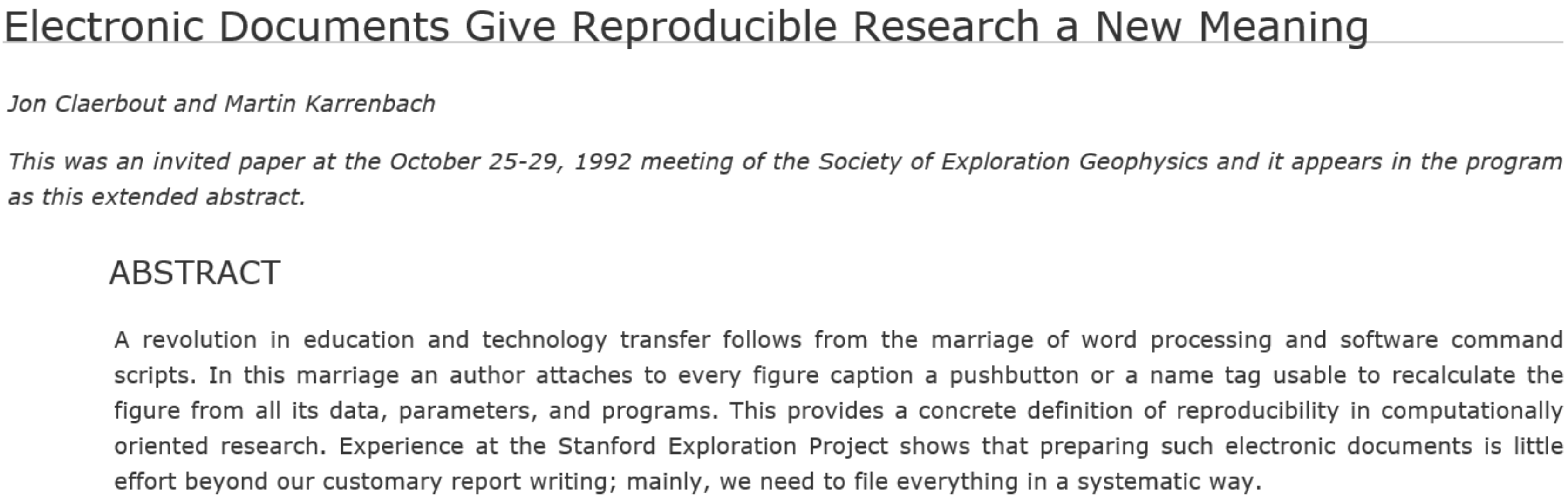
ORCID - Open Research & Contributor ID

- Aims to solve the author/contributor name ambiguity problem in scholarly communications
- Central registry of unique identifiers for individual researchers
- Open and transparent linking mechanism between ORCID and other current author ID schemes.
- Identifiers can be linked to the researcher's output to enhance the scientific discovery process

Research Reproducibility and Computational Science

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.

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
  - Data needed for others to reproduce the results
- Code and methods must be available and accessible
  - Computer scripts and workflow pipelines
- Citation
  - DOIs for both data and software
- Copyright and publisher agreements
  - Need only give publisher permission to publish – as for US Government
- Supplemental materials
  - Need to follow best practices

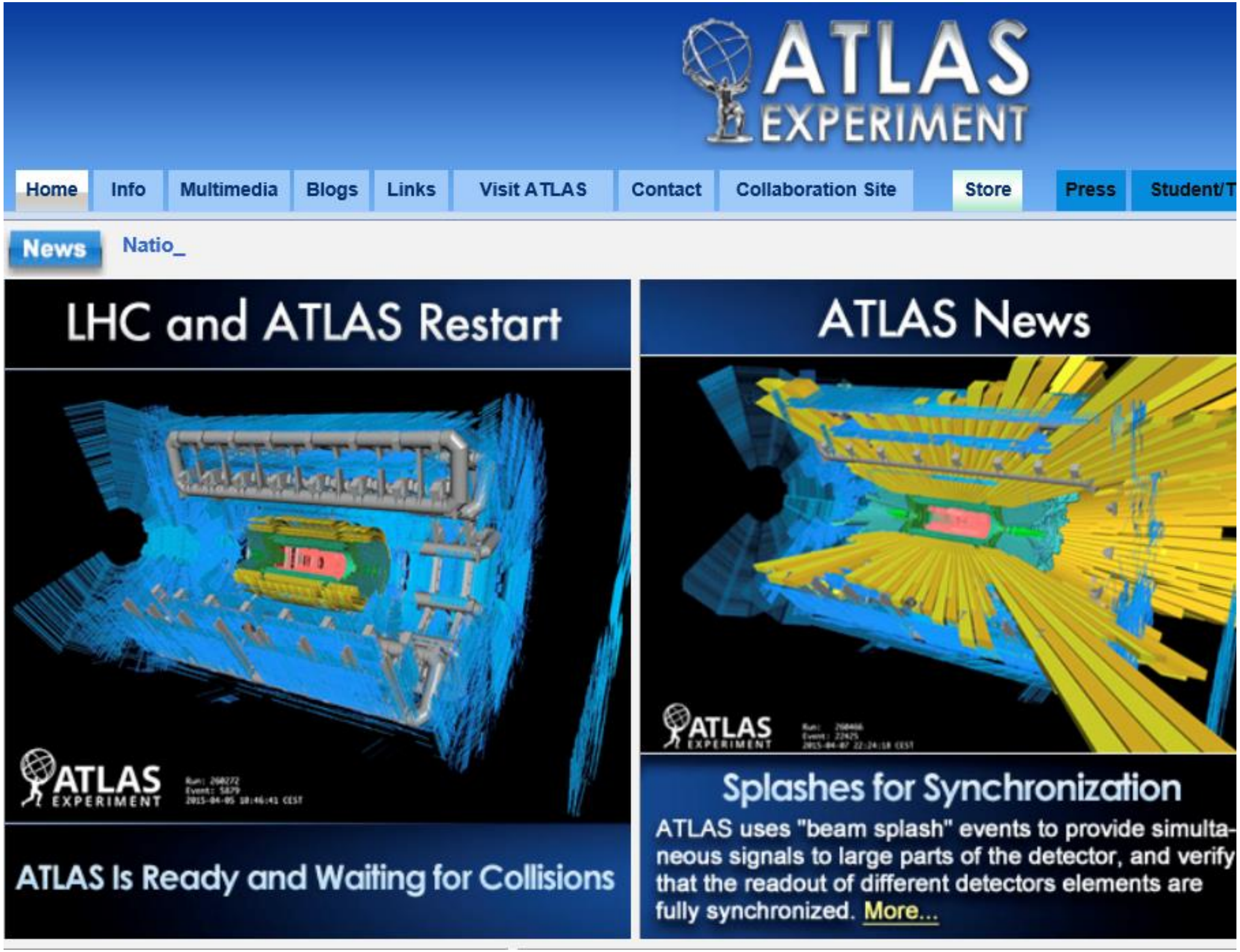
From **http://wiki.stodden.net**

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

Research Reproducibility at the LHC



LHC and ATLAS Restart

ATLAS Is Ready and Waiting for Collisions

ATLAS News

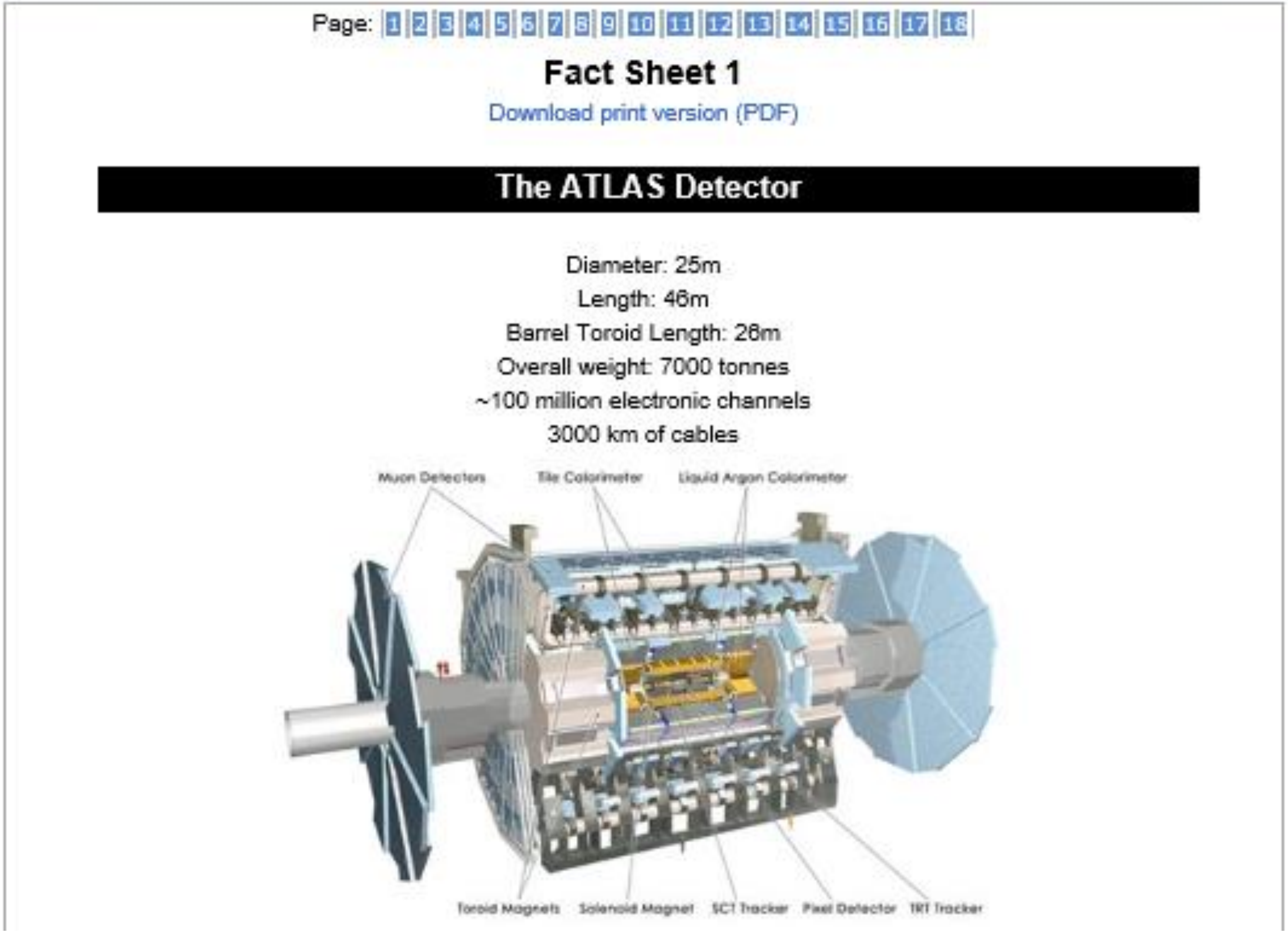
Splashes for Synchronization

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

- 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 ☺

Thanks to Gordon Watts, UW



Page: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Fact Sheet 1

Download print version (PDF)

The ATLAS Detector

Diameter: 25m

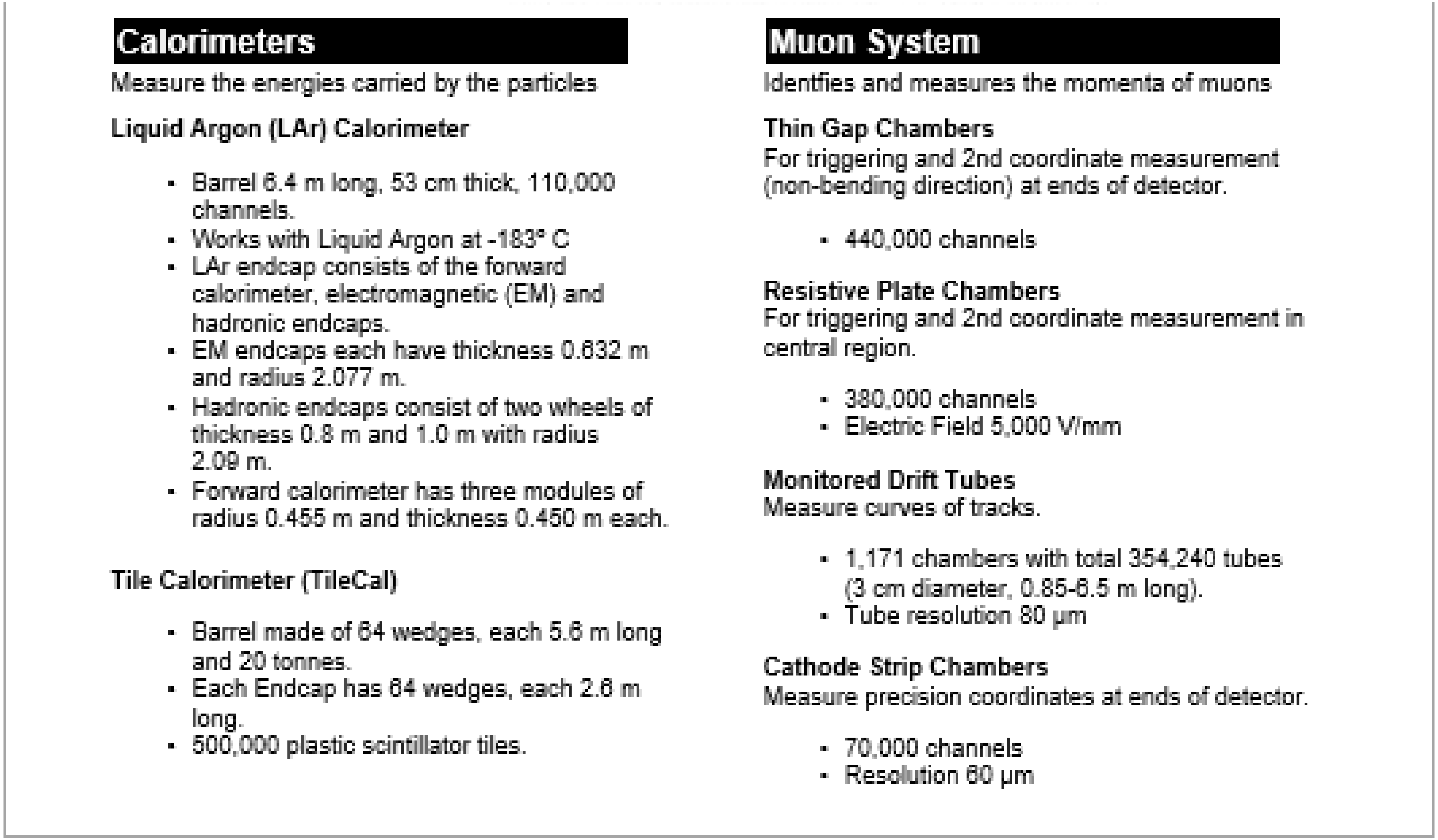
Length: 46m

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° 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 2.6 m long.
- 500,000 plastic scintillator tiles.

Muon System

Identfies 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 µm

Cathode Strip Chambers

Measure precision coordinates at ends of detector.

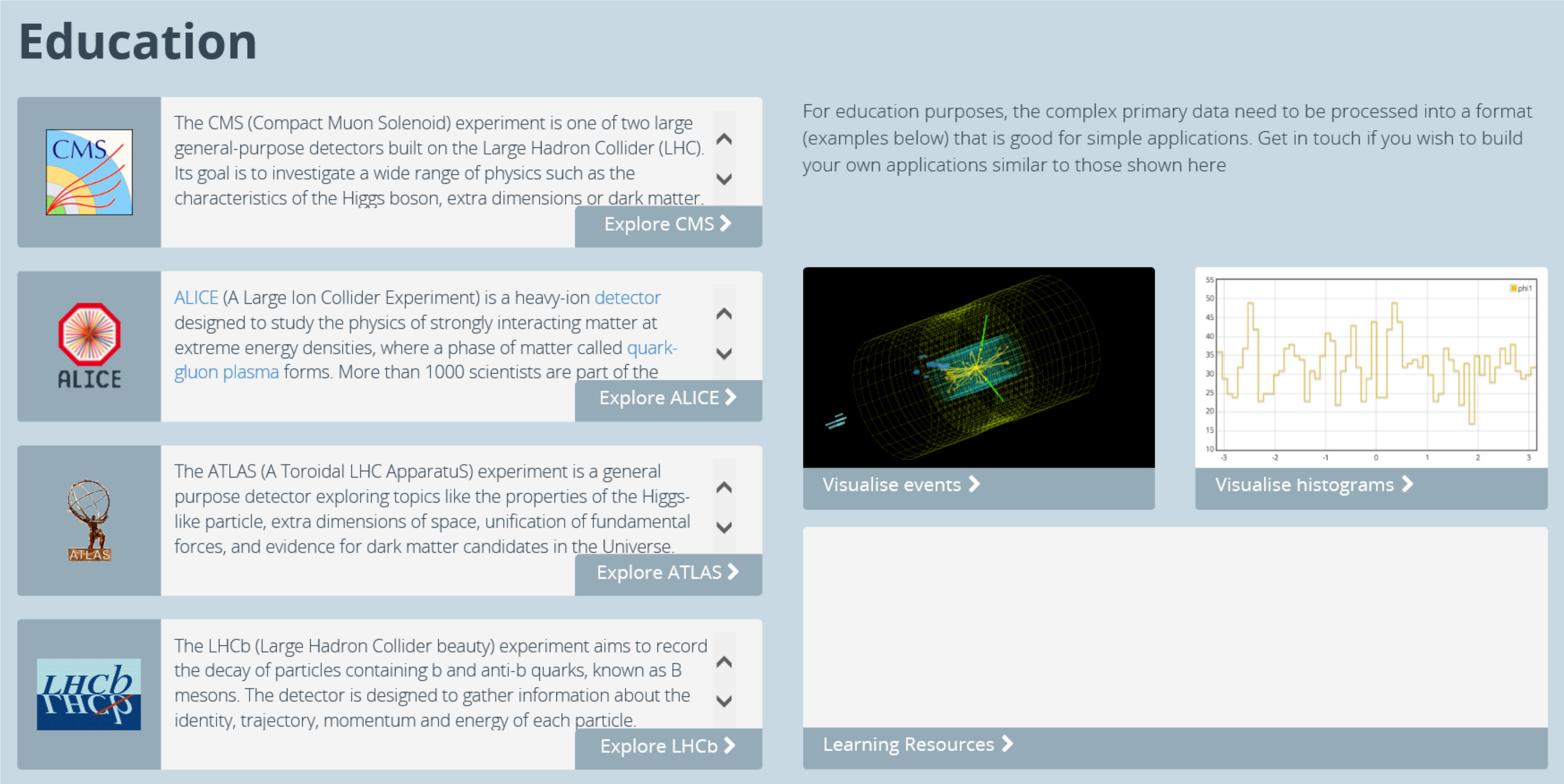
- 70,000 channels
- Resolution 60 µm

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
- Only after observed differences have been investigated and resolved, are new versions of the code released to whole ATLAS collaboration
- ATLAS uses Apache Subversion (SVN) version-control system
- With over 2000 software packages to be tracked, ATLAS developed its own release management software

Research Reproducibility?

- 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
    - See the Higgs boson at the same mass value in both experiments
- Making meaningful data available to the public is difficult
    - 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

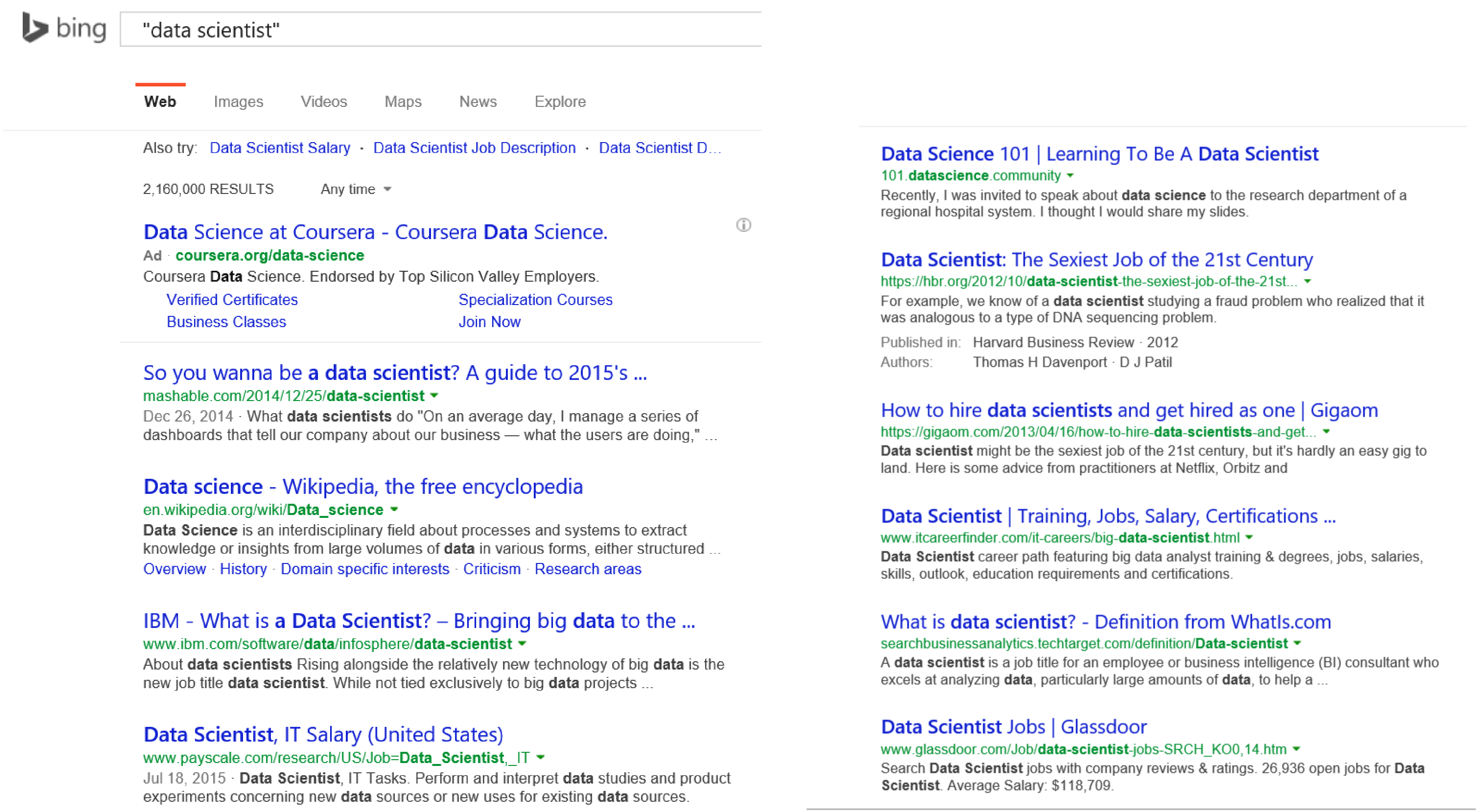
Visualise events ›

Visualise histograms ›

Learning Resources ›

http://opendata.cern.ch

Data Scientists in the Future?



"data scientist"

Also try: Data Scientist Salary · Data Scientist Job Description · Data Scientist D…

2,160,000 RESULTS

Data Science at Coursera - Coursera **Data** Science.
Ad · **coursera.org/data-science**
Coursera **Data** Science. Endorsed by Top Silicon Valley Employers.
Verified Certificates · Specialization Courses
Business Classes · Join Now

So you wanna be **a data scientist**? A guide to 2015's ...
mashable.com/2014/12/25/**data-scientist**
Dec 26, 2014 · What **data scientists** do "On an average day, I manage a series of dashboards that tell our company about our business — what the users are doing," …

Data science - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Data_science**
Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of **data** in various forms, either structured …
Overview · History · Domain specific interests · Criticism · Research areas

IBM - What is **a Data Scientist**? – Bringing big **data** to the ...
www.ibm.com/software/**data**/infosphere/**data-scientist**
About **data scientists** Rising alongside the relatively new technology of big **data** is the new job title **data scientist**. While not tied exclusively to big **data** projects ...

Data Scientist, IT Salary (United States)
www.payscale.com/research/US/Job=**Data_Scientist**,_IT
Jul 18, 2015 · **Data Scientist**, IT Tasks. Perform and interpret **data** studies and product experiments concerning new **data** sources or new uses for existing **data** sources.

Data Science 101 | Learning To Be A **Data Scientist**
101.**datascience**.community
Recently, I was invited to speak about **data science** to the research department of a regional hospital system. I thought I would share my slides.

Data Scientist: The Sexiest Job of the 21st Century
https://hbr.org/2012/10/**data-scientist**-the-sexiest-job-of-the-21st...
For example, we know of a **data scientist** studying a fraud problem who realized that it was analogous to a type of DNA sequencing problem.
Published in: Harvard Business Review · 2012
Authors: Thomas H Davenport · D J Patil

How to hire **data scientists** and get hired as one | Gigaom
https://gigaom.com/2013/04/16/how-to-hire-**data-scientists**-and-get...
Data scientist might be the sexiest job of the 21st century, but it's hardly an easy gig to land. Here is some advice from practitioners at Netflix, Orbitz and

Data Scientist | Training, Jobs, Salary, Certifications ...
www.itcareerfinder.com/it-careers/big-**data-scientist**.html
Data Scientist career path featuring big data analyst training & degrees, jobs, salaries, skills, outlook, education requirements and certifications.

What is **data scientist**? - Definition from WhatIs.com
searchbusinessanalytics.techtarget.com/definition/**Data-scientist**
A **data scientist** is a job title for an employee or business intelligence (BI) consultant who excels at analyzing **data**, particularly large amounts of **data**, to help a ...

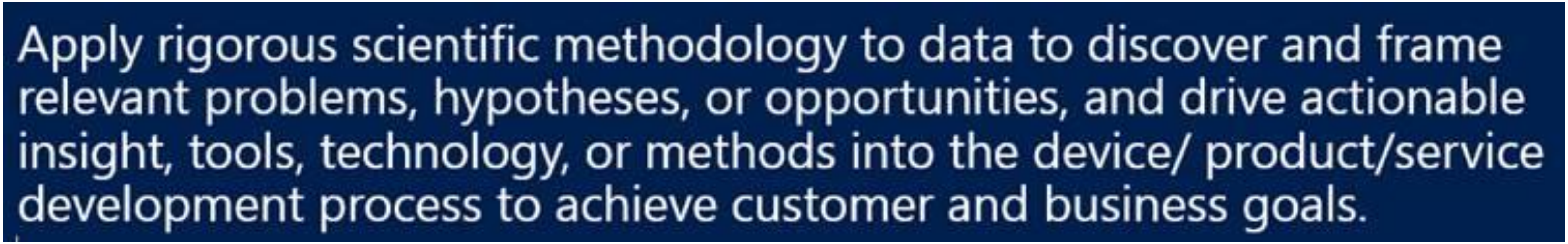
Data Scientist Jobs | Glassdoor
www.glassdoor.com/Job/**data-scientist**-jobs-SRCH_KO0,14.htm
Search **Data Scientist** jobs with company reviews & ratings. 26,936 open jobs for **Data Scientist**. Average Salary: $118,709.

Microsoft – advertizing for Data Scientists

DATA & APPLIED SCIENTIST

3 ROLES:

- DATA SCIENTIST
- MACHINE LEARNING SCIENTIST
- APPLIED SCIENTIST



Apply rigorous scientific methodology to data to discover and frame relevant problems, hypotheses, or opportunities, and drive actionable insight, tools, technology, or methods into the device/ product/service development process to achieve customer and business goals.

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

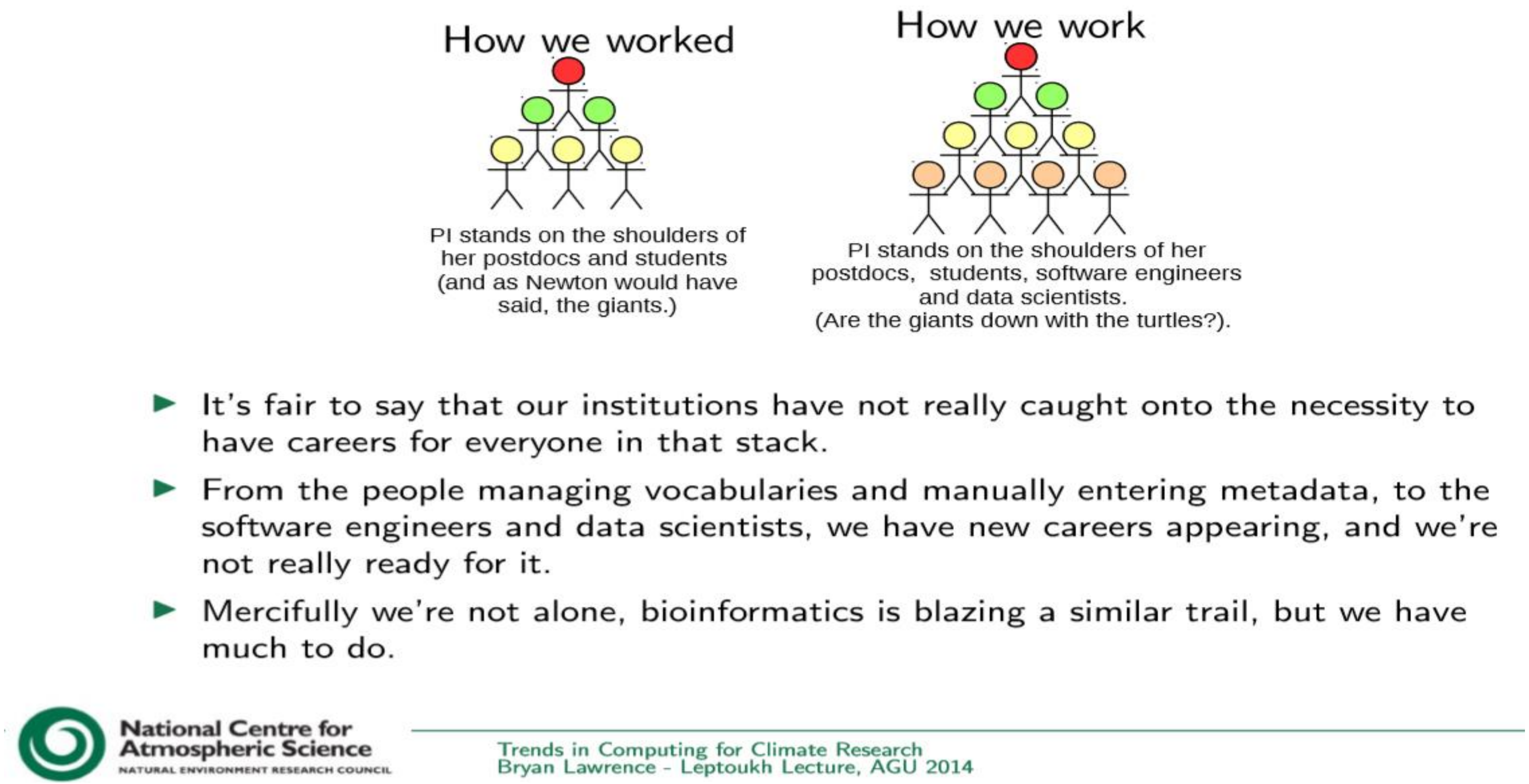
CONCEPTUALISE
CREATE OR RECEIVE
TRANSFORM
DISPOSE
CURATE
APPRAISE & SELECT
ACCESS, USE & REUSE
PRESERVATION PLANNING
DESCRIPTION
REPRESENTATION INFORMATION
COMMUNITY WATCH & PARTICIPATION
STORE
INGEST
PRESERVE
PRESERVATION ACTION
MIGRATE
REAPPRAISE

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

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.

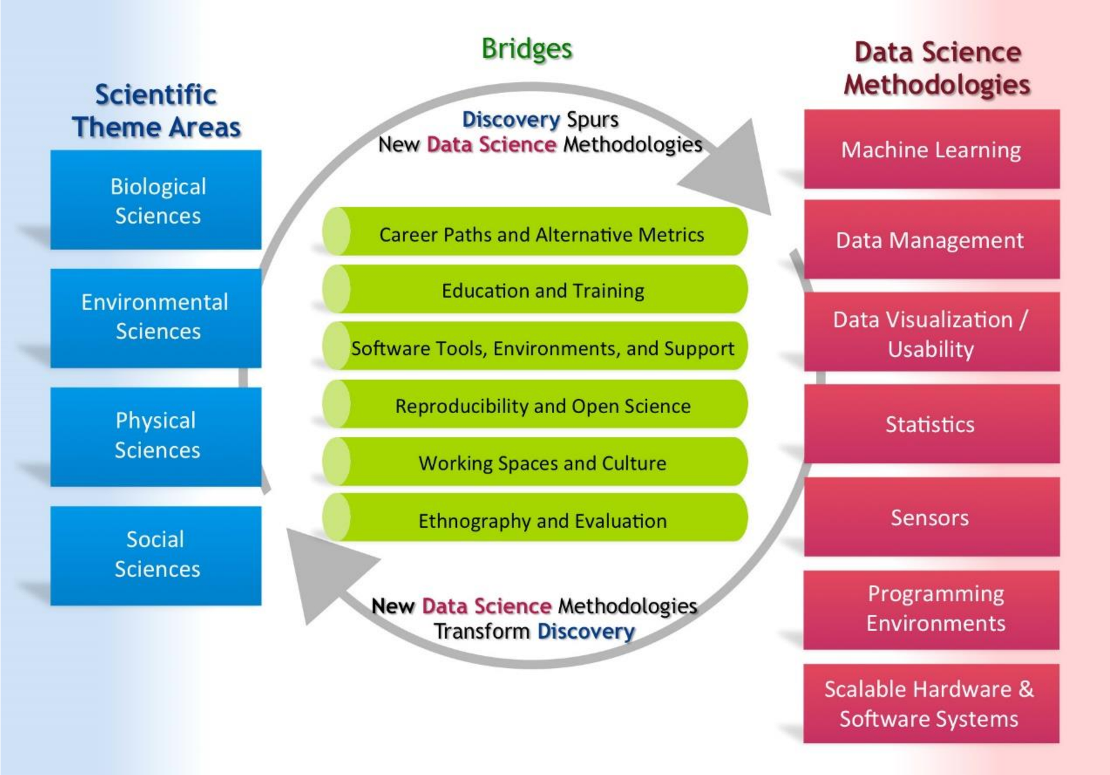
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Trends in Computing for Climate Research
Bryan Lawrence - Leptoukh Lecture, AGU 2014

Slide thanks to Bryan Lawrence

The Moore/Sloan Data Science Environments Project

Launched late fall 2013



Bridges

Scientific Theme Areas

- Biological Sciences
- Environmental Sciences
- Physical Sciences
- Social Sciences

Discovery Spurs New Data Science Methodologies

- Career Paths and Alternative Metrics
- Education and Training
- Software Tools, Environments, and Support
- Reproducibility and Open Science
- Working Spaces and Culture
- Ethnography and Evaluation

New Data Science Methodologies Transform Discovery

Data Science Methodologies

- Machine Learning
- Data Management
- Data Visualization / Usability
- Statistics
- Sensors
- Programming Environments
- Scalable Hardware & Software Systems

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

NYU

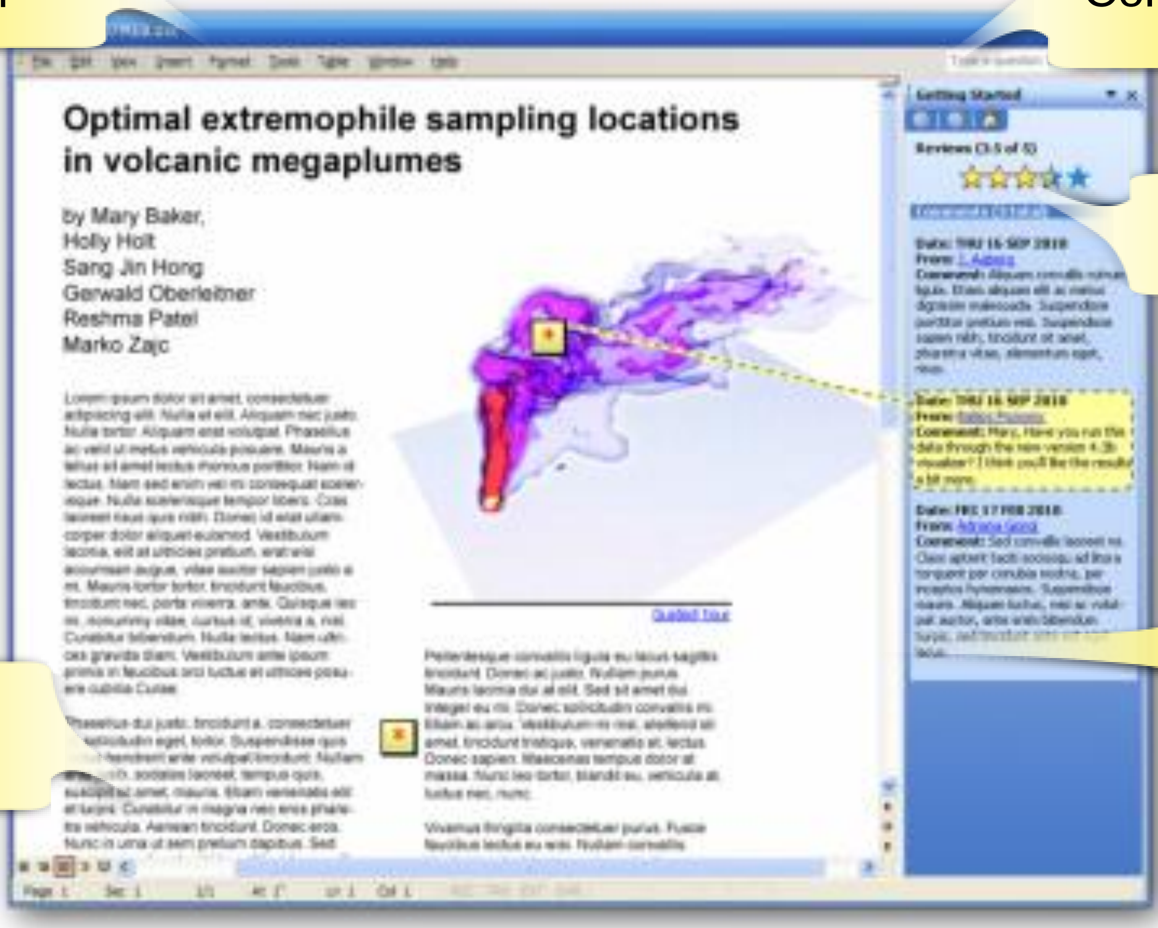
Berkeley UNIVERSITY OF CALIFORNIA

W UNIVERSITY of WASHINGTON

Slide with thanks to Ed Lazowska, UW eScience Institute

Open Science

Vision for a New Era of Research Reporting



Optimal extremophile sampling locations in volcanic megaplumes

by Mary Baker,
Holly Holt
Sang Jin Hong
Gerwald Oberleitner
Reshma Patel
Marko Zajc

Reproducible Research

Collaboration

Reputation & Influence

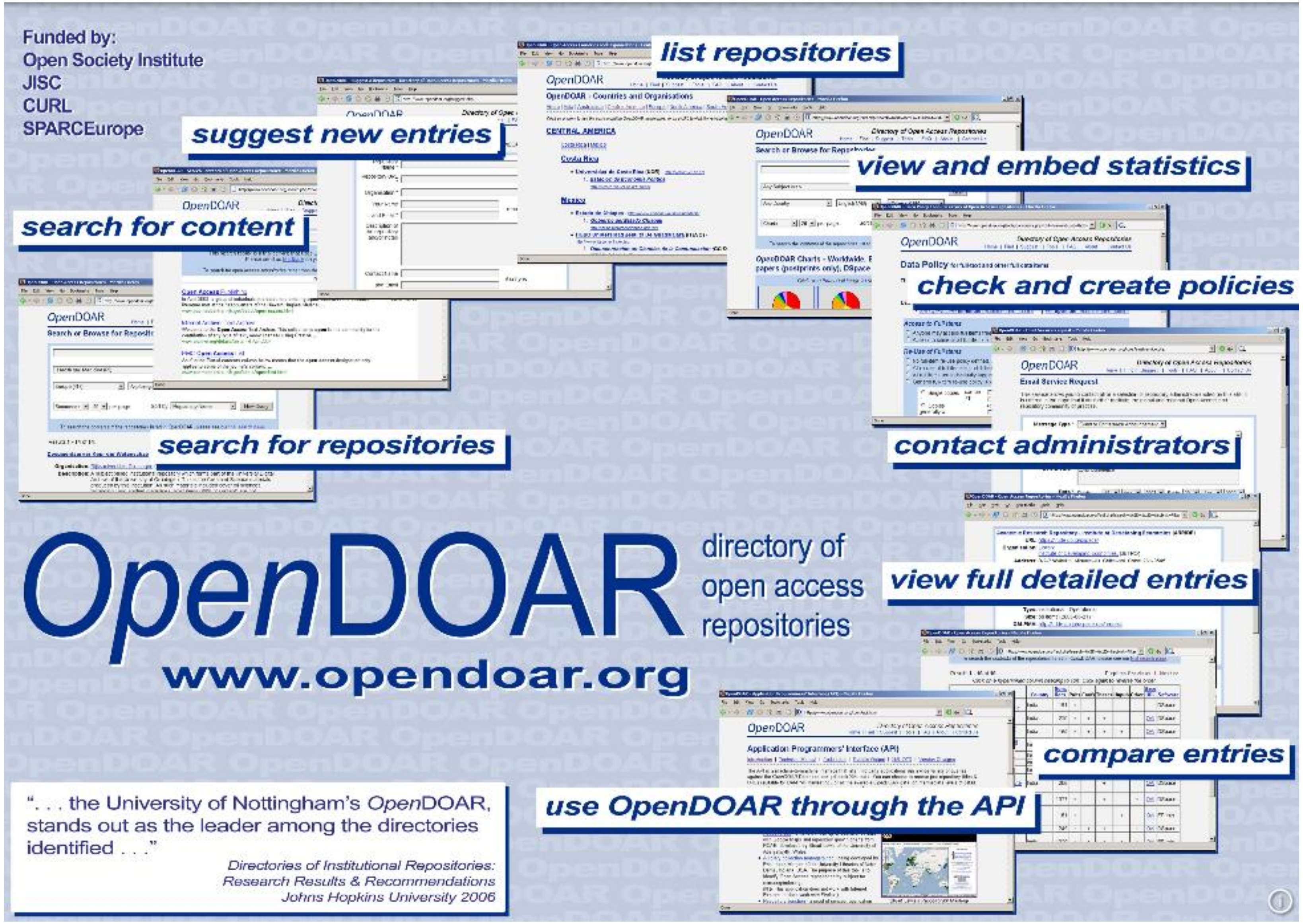
Dynamic Documents

Interactive Data

(Thanks to Bill Gates SC05)

Role of Research Libraries?

Institutional Research Repositories



Funded by:
Open Society Institute
JISC
CURL
SPARCEurope

list repositories

suggest new entries

view and embed statistics

search for content

check and create policies

search for repositories

contact administrators

OpenDOAR

directory of
open access
repositories

www.opendoar.org

view full detailed entries

compare entries

use OpenDOAR through the API

"... the University of Nottingham's OpenDOAR, stands out as the leader among the directories identified ..."

Directories of Institutional Repositories:
Research Results & Recommendations
Johns Hopkins University 2006

UK Funders Expectations for Data Preservation

- Research organisations will ensure that 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

Progress in Data Curation in last 10 years?

- Biggest change is funding agency mandate:
  - NSF’s insistence on a 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 the quality fraction is increasing!
- Frew’s laws of metadata:
  - First law: scientists don’t write metadata
  - 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 a.k.a. science librarians

With thanks to Jim Frew, UCSB

Three final comments on Open Science

Paul Ginsparg, creator of arXiv, on the open access revolution:

‘Ironically, it is also possible that the technology of the 21st century will allow the traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research Enterprise.’

Someone praising Helen Berman, Head of the Protein Data Bank PDB:

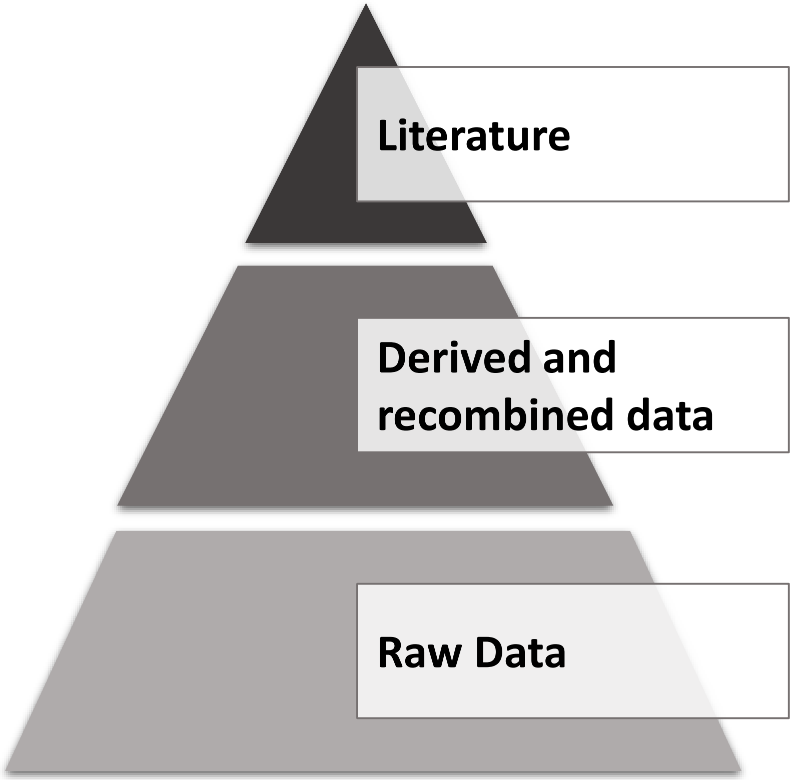
‘One of the remarkable things about Helen is that her life has been devoted to service within science rather than, as some might call it, doing real science.’

Michael Lesk on Just-in-time instead of Just-in-case?

‘Most of the cost of archiving is spent at the start, before we know whether the articles will be read or the data used. With data, with no emotional investment in peer review, it might be easier to do a simpler form of deposit, where as much as possible is postponed till the data are called for. ‘

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)