# The Fourth Paradigm: Data-Intensive Scientific Discovery, Open Science and the Cloud

**Tony Hey**

**Senior Data Science Fellow**

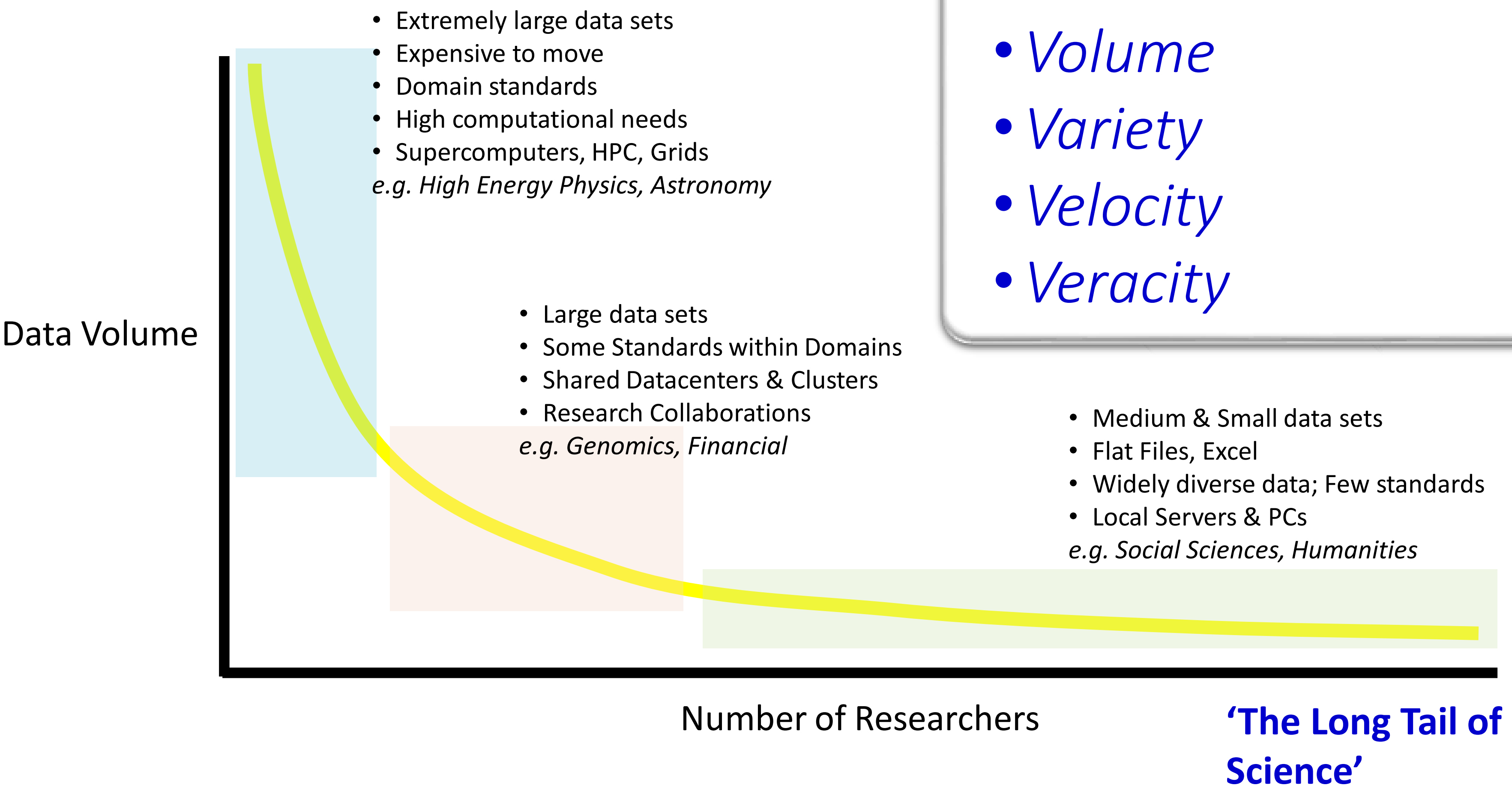
**eScience Institute**

**University of Washington**

tony.hey@live.com

# 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

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

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

*Sloan Foundation, NSF, DOE, NASA*

# 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 new data from the SDSS Data Release 1 (DR1).

More...

For Astronomers

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

More...

SkyServer Tools

- Famous places
- Get images
- Sending sky
- Visual Tools
- Explore
- Search
- Object upload

Science Projects

- Basic
- Advanced
- Challenges
- For Kids
- Games and Contests
- Teachers
- Links to other projects

Info Links

- About Astronomy
- About the SDSS
- About the SkyServer
- The EDR SkyServer
- SDSS Data Release 1
- SkyQuery
- Download

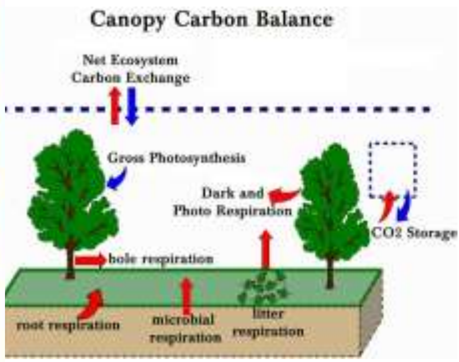
Help

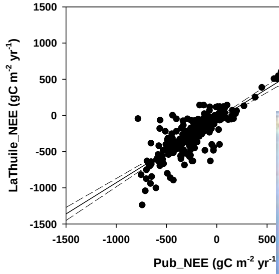
- Getting Started
- FAQ
- How To
- Glossary
- Schema Browser
- Introduction to SQL
- Expert Background

# Carbo-Climate Synthesis

- Role of photosynthesis in global warming?
  - Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict
  - Difference is either due to plants or ocean absorption.
- Communal field science – each investigator acts independently.
  - Cross site studies and integration with modeling increasingly important
- Sharepoint site:

  [www.fluxdata.org](http://www.fluxdata.org)
  - 921 site-years of data from 240 sites around the world; 80+ site-years now being added
  - 60+ paper writing teams
  - American data subset is public and served more widely
  - Summary data products greatly simplify initial data discovery




LaThuile_NEE (gC m$^{-2}$ yr$^{-1}$)
1500
1000
500
0
-500
-1000
-1500
-1500
-1000
-500
0
500
Pub_NEE (gC m$^{-2}$ yr$^{-1}$)

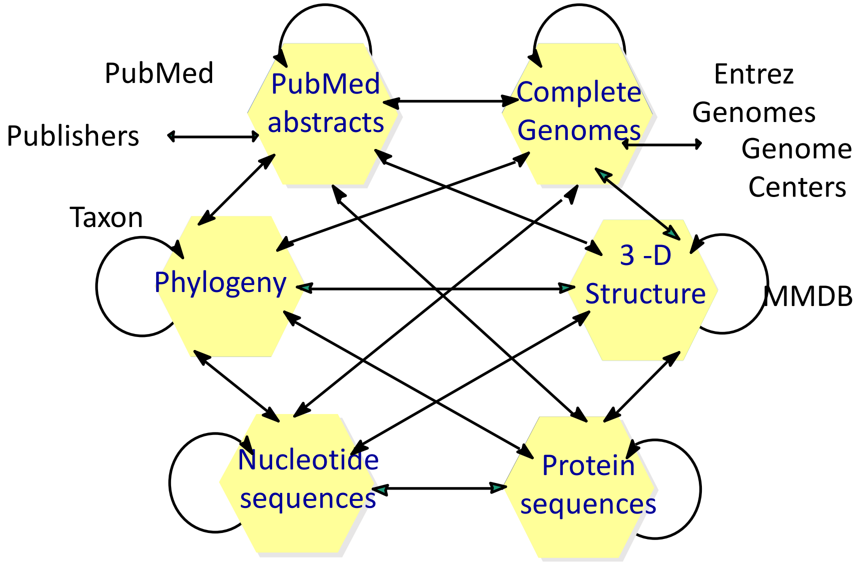
(Dennis Baldocchi (Berkeley Water Center)

And Catharine van Ingen (Microsoft Research))

# 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



PubMed

Publishers

Taxon

PubMed abstracts

Complete Genomes

Entrez Genomes

Genome Centers

Phylogeny

3 -D Structure

MMDB

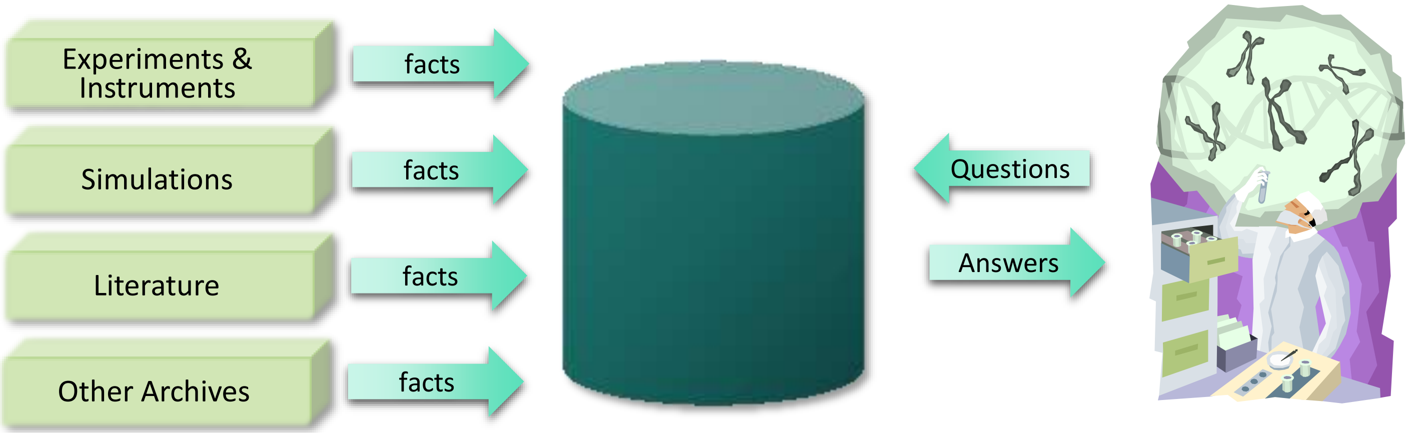
Nucleotide sequences

Protein sequences

**Entrez cross-database search**

# X-Info

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



Experiments & Instruments

Simulations

Literature

Other Archives

facts

facts

facts

facts

Questions

Answers

## The Generic Problems

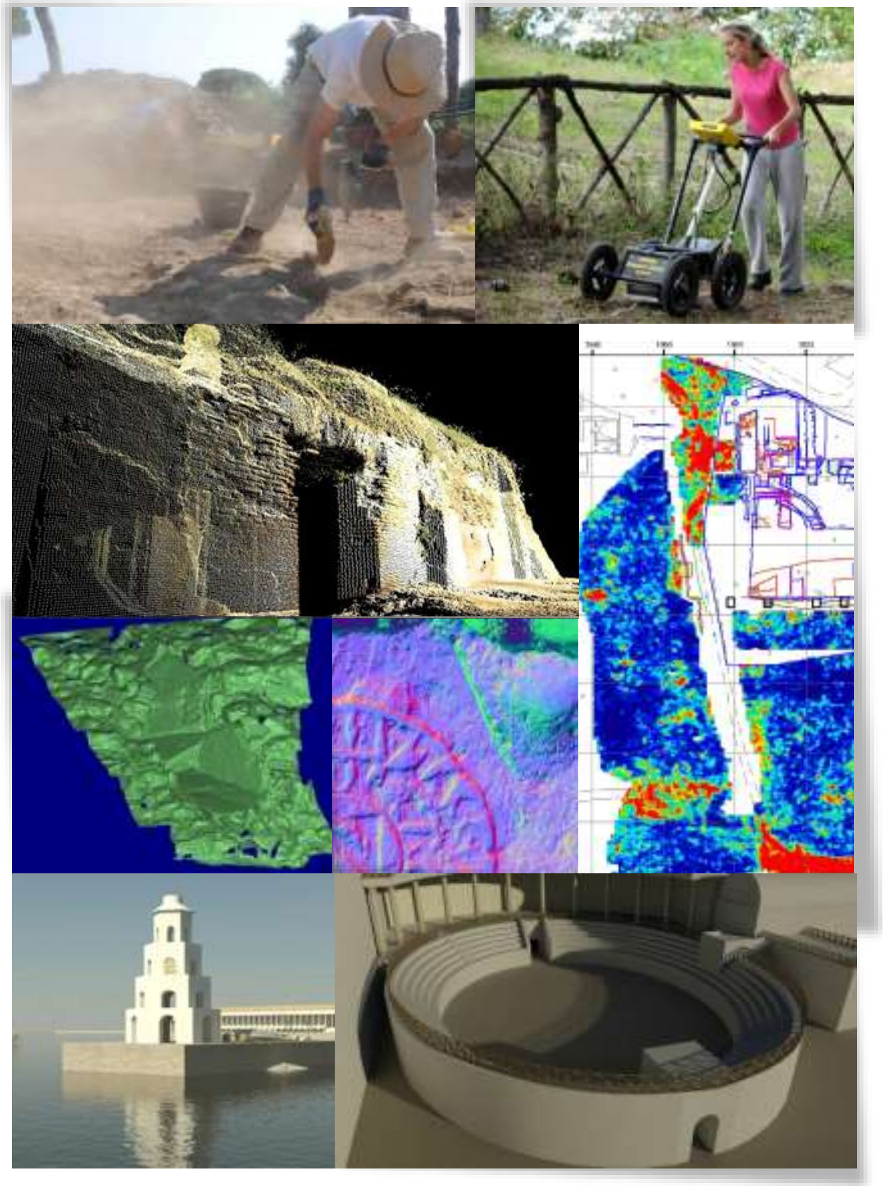
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

**Slide thanks to Jim Gray**

Keynote by Dan Fay, director of E3 at Microsoft Research Connections, on “The Rise of X-Informatics.

# Archaeo-Informatics

- Archaeologists must capture and organize artifacts and data
- Multiple sources, from excavation and hand sifting, to advanced geophysics and aerial surveying
- Context is everything, ultimately visualize and synthesize
- Advanced computational tools from data management and processing, to analysis and visualization
- Example is AHRC Portus Project examining early port of Rome



**PI: Graeme Earl**
**University of Southampton**

# 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

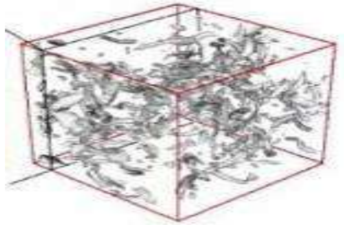
Today – **Data-Intensive Science**

- 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

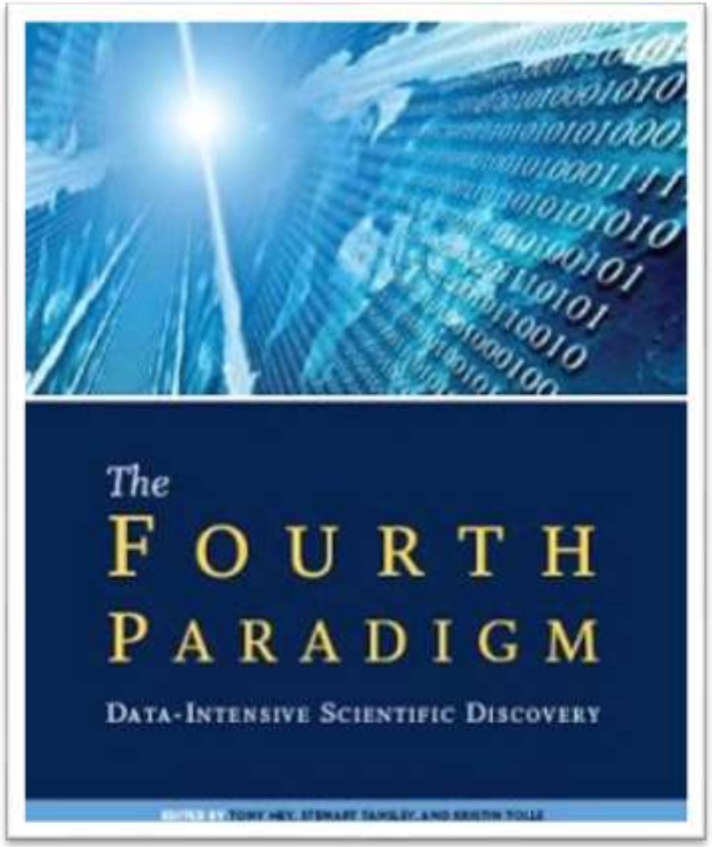
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

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



*(With thanks to Jim Gray)*

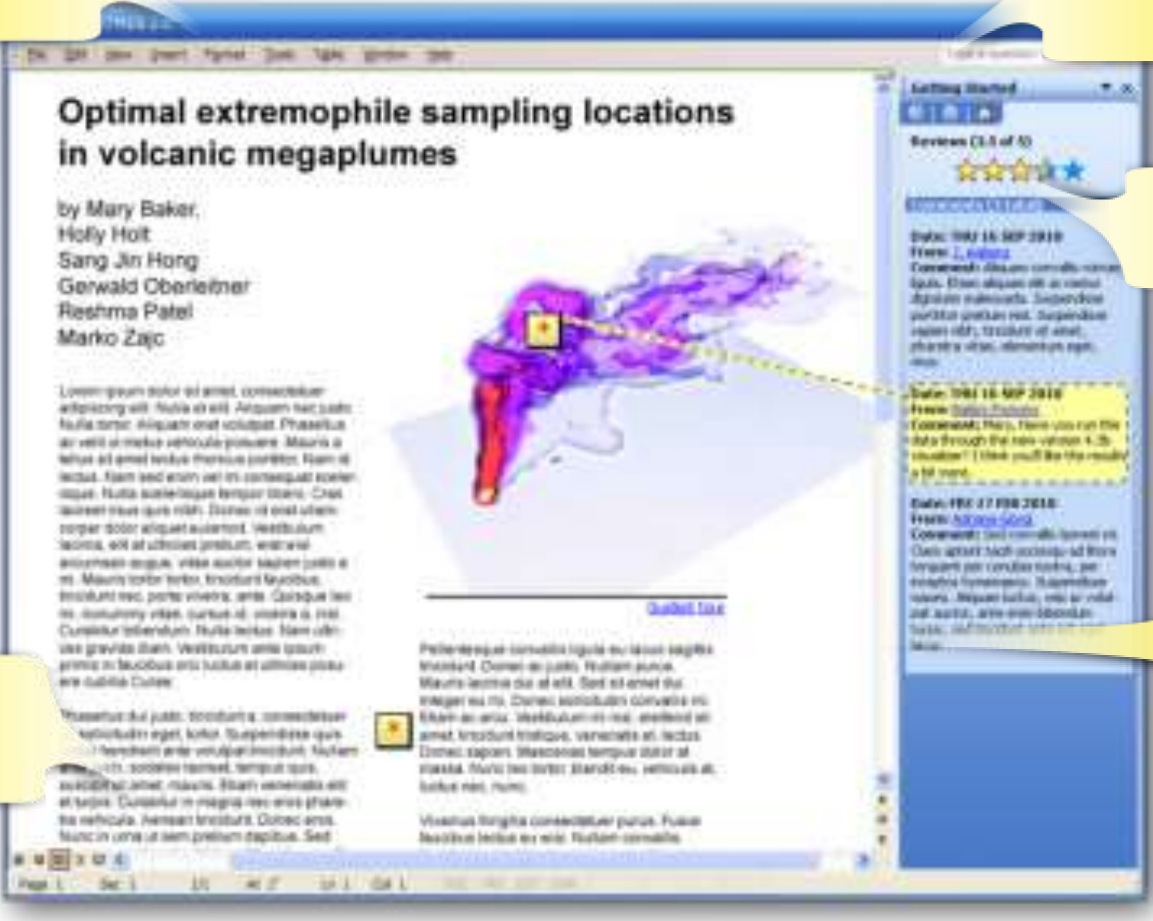
# eScience and Data-Intensive Scientific Discovery


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**

# Open Access, Open Data, 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

Comment: Mary, Have you run the data through the new version 4.3b visualizer? I think you'll like the results a bit more.

Reproducible Research

Collaboration

Reputation & Influence

Interactive Data

Dynamic Documents

*(Thanks to Bill Gates SC05)*

# US White House 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

# 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 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

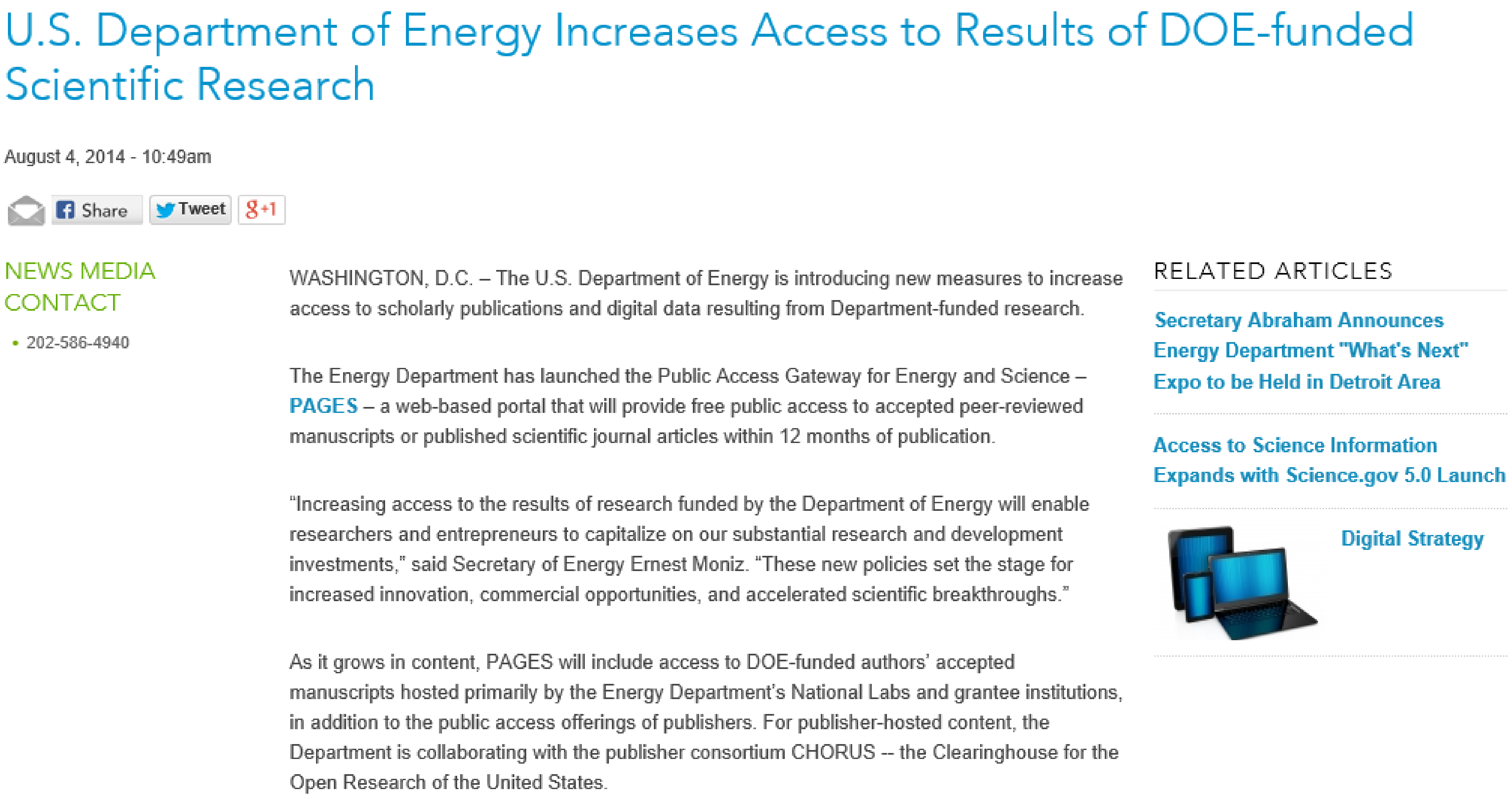
2 August 2013

# 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%



# 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.

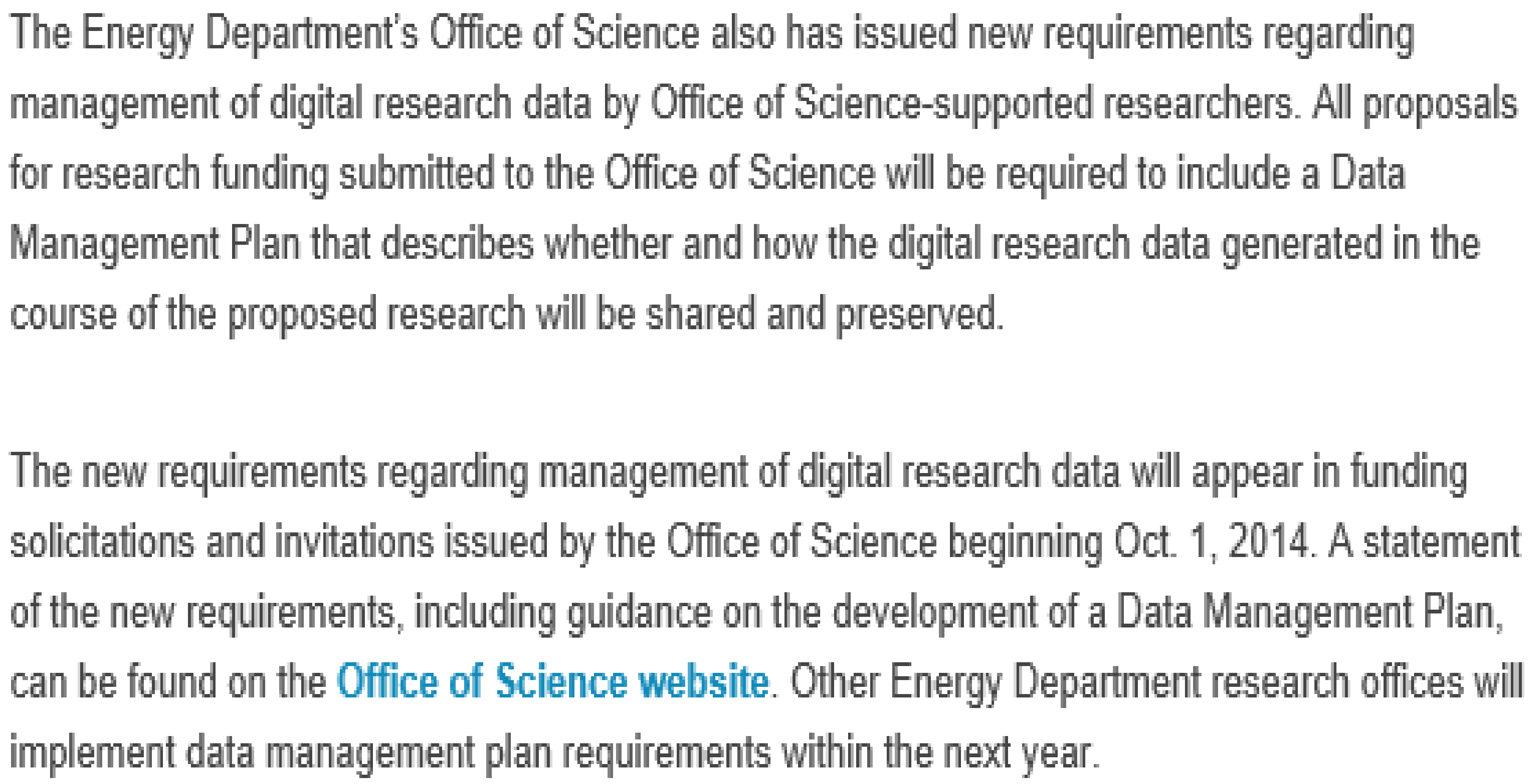
RELATED ARTICLES

**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

# 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

# Data Curation:
# State of the Art

# Long Term Access to Large Scientific Data Sets: The SkyServer and Beyond

NSF ACI Data Infrastructure Building Blocks (DIBBS) Program

- $7.6M project
- Started 2013 – end date 2018

Project goals

- Address curation issues arising from the data and service life-cycle
- Support small but complex data in the 'Long Tail' of science.

➢ Need to curate both Data and Services lifecycle

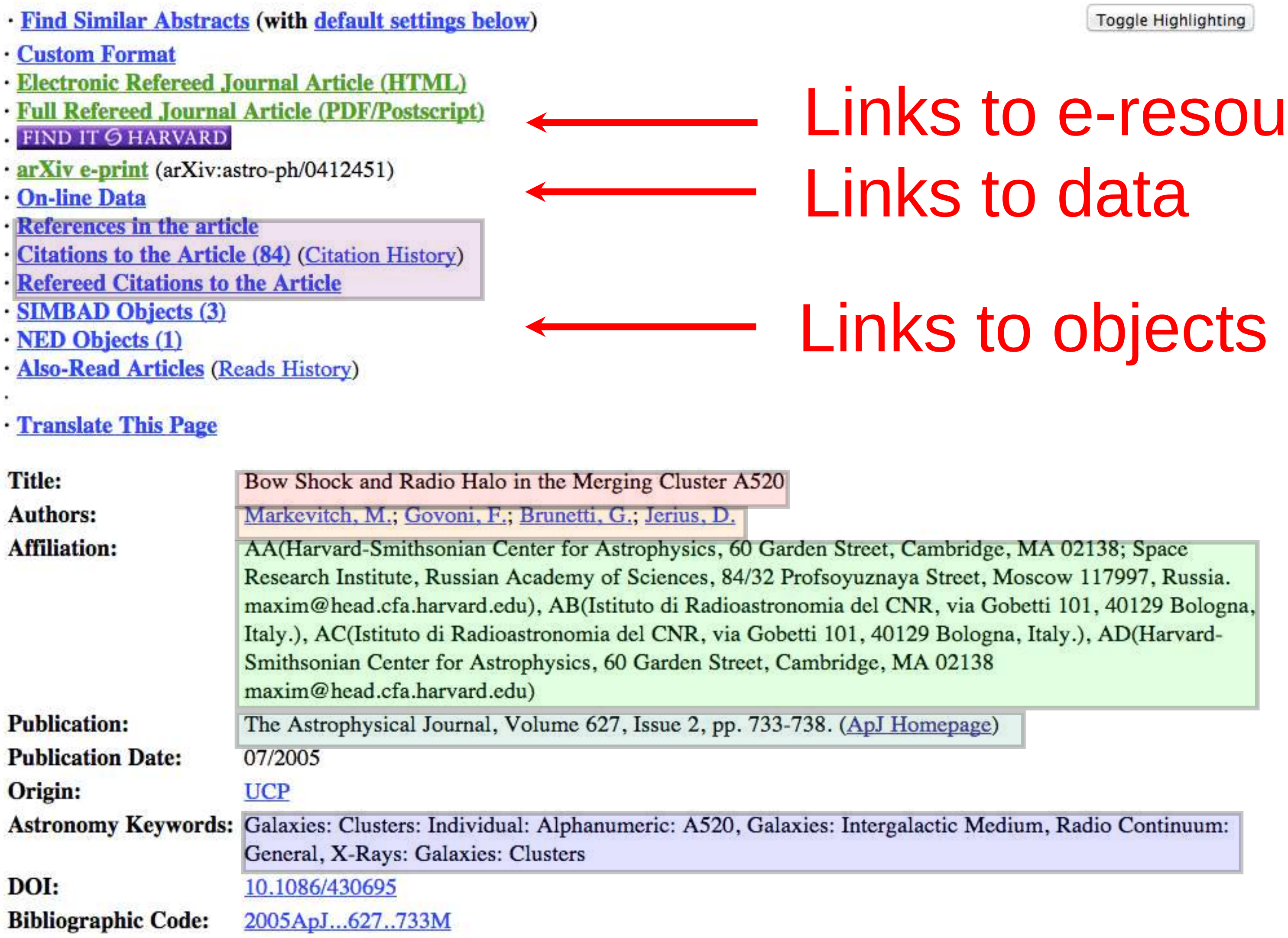
# What happened to Virtual Observatories?

- UK AstroGrid project
  - Funding cancelled in 2008

- US Virtual Astronomy Observatory (VAO)
  - Project funding discontinued in 2014

➢ But much of the infrastructure, tools and technology still lives on with participation in the International Virtual Observatory Alliance (IVOA)

# 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

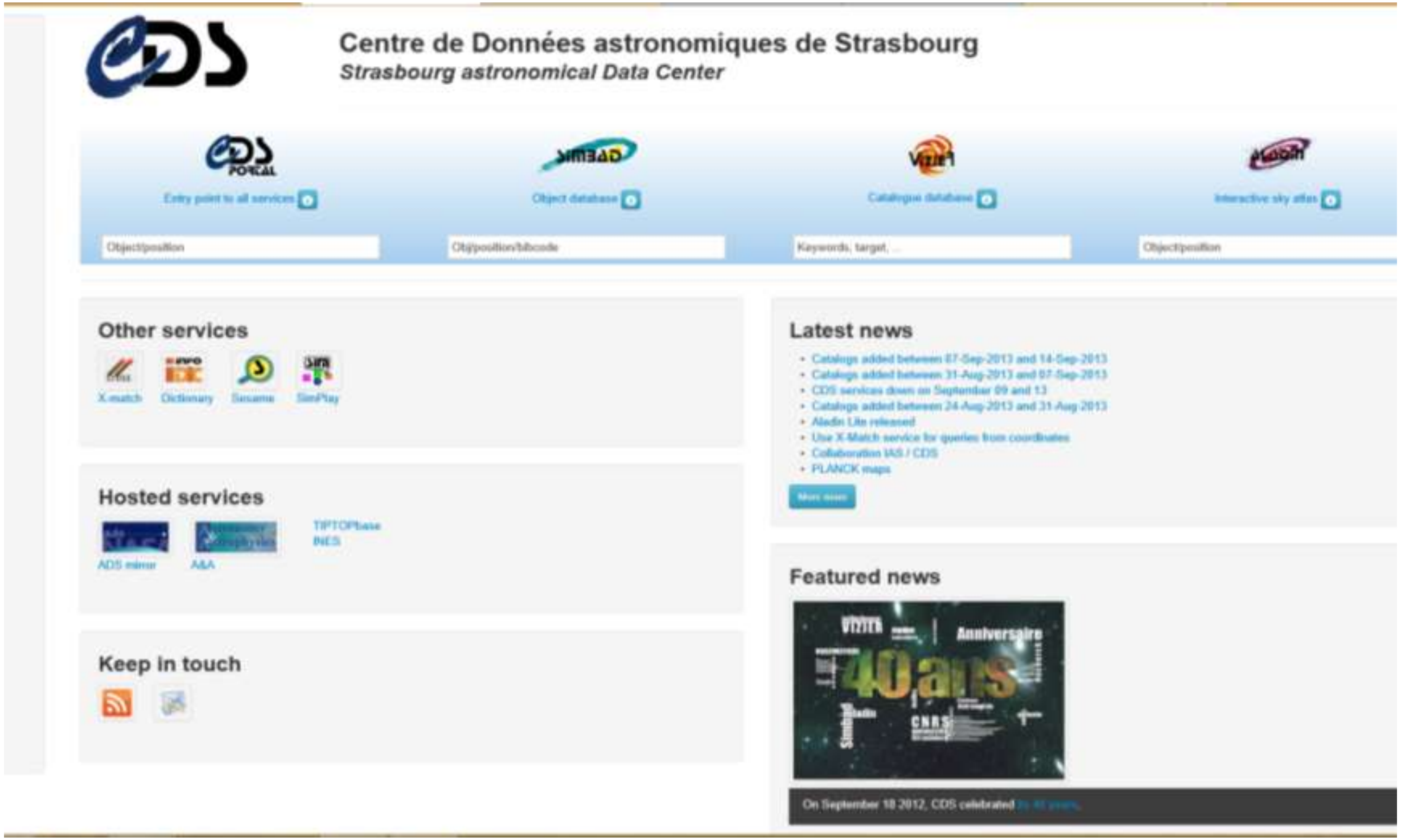
| | |
|---|---|
| **Title:** | Bow Shock and Radio Halo in the Merging Cluster A520 |
| **Authors:** | Markevitch, M.; Govoni, F.; Brunetti, G.; Jerius, D. |
| **Affiliation:** | AA(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; Space Research Institute, Russian Academy 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

# Strasbourg CDS Datasets



Centre de Données astronomiques de Strasbourg

*Strasbourg astronomical Data Center*

## Other services

## Hosted services

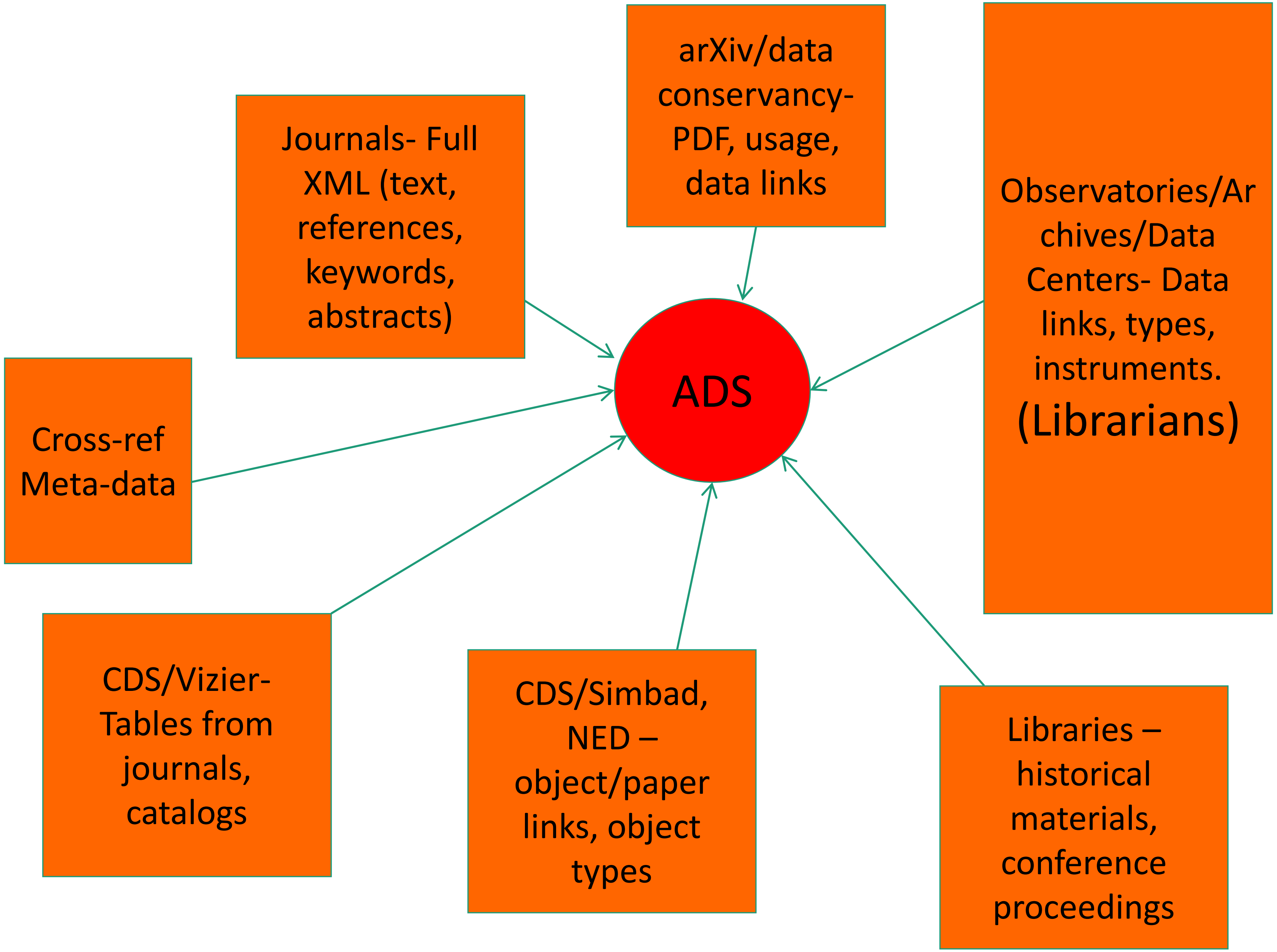
## Keep in touch

## 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

## Featured news

On September 18 2012, CDS celebrated its 40 years.



Journals- Full XML (text, references, keywords, abstracts)

arXiv/data conservancy- PDF, usage, data links

Observatories/Archives/Data Centers- Data links, types, instruments.

(Librarians)

Cross-ref Meta-data

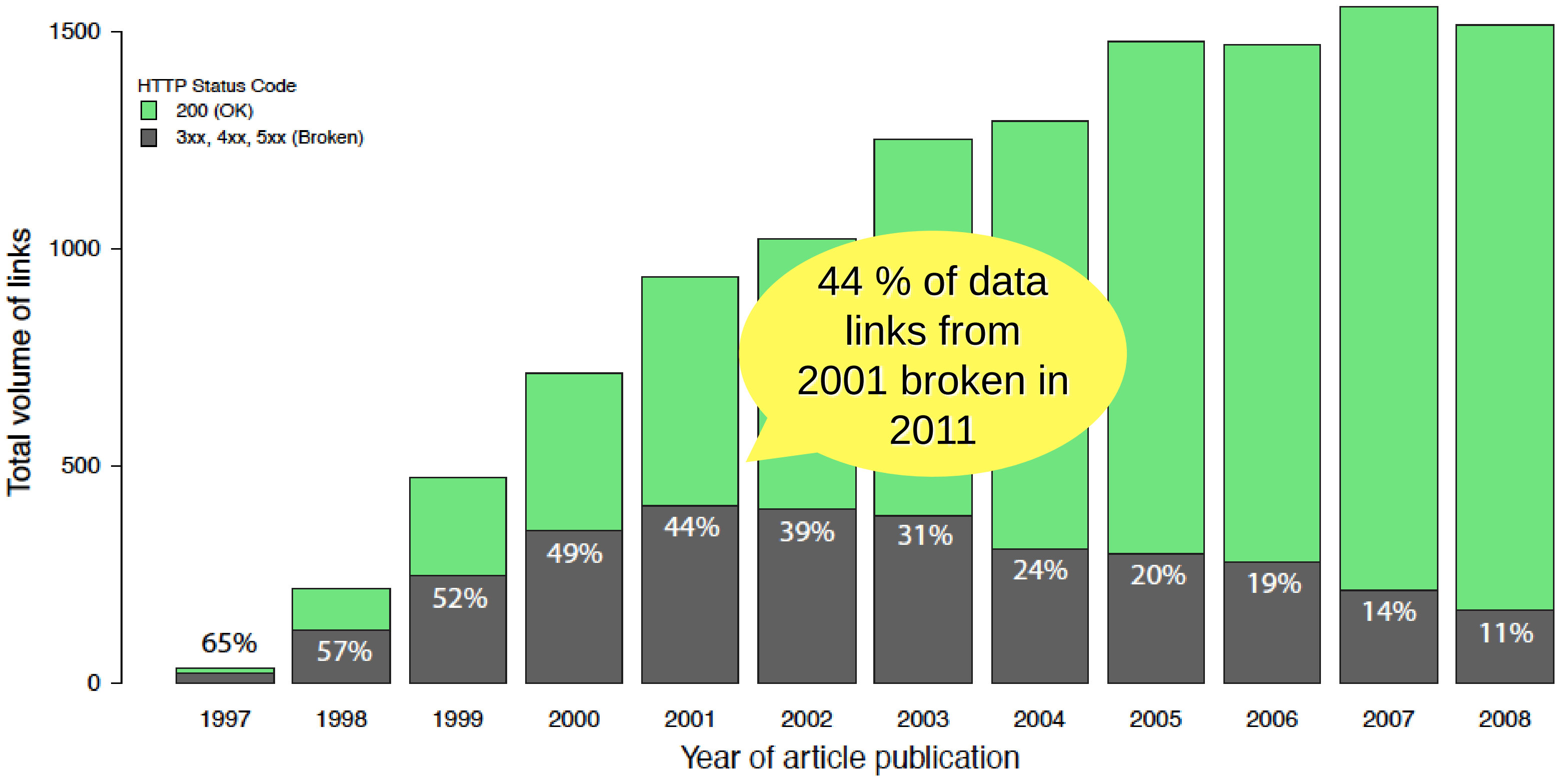
ADS

CDS/Vizier- Tables from journals, catalogs

CDS/Simbad, NED – object/paper links, object types

Libraries – historical materials, conference proceedings

# But 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*

# Progress since 2004?
# The View from ADS (Michael Kurtz)

Comments:

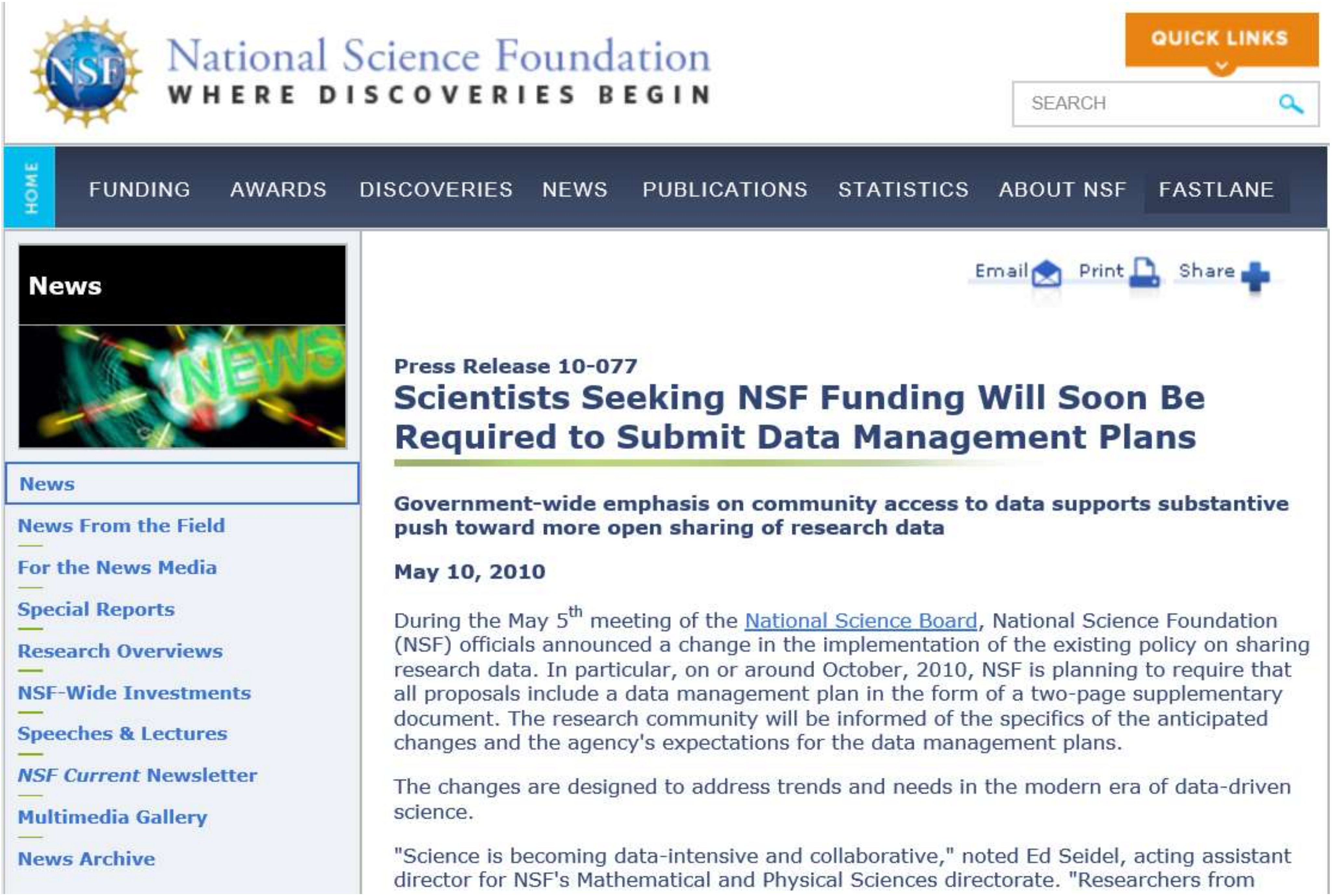
- Does not see much progress in the last ten years: now one step back, waiting for the next two steps forward
- Ten years ago concerned that the Virtual Observatory would suffocate itself with bureaucracy: unfortunately this has now happened …
- New large repositories (Zenodo, Dataverse) are creating an infrastructure of almost entirely uncurated data

➢The problem with curation is that the funding is almost entirely local but in the digital world the use is mainly global. Leads to tragedy of the commons where no one will assume long-term obligation to curate and manage data which is mainly not from local sources.

# Progress since 2004?
# The View from CDS (Francoise Genova)

There are two major areas of progress:

- VO Framework
  - Interoperability framework with aspects on data description, formats, vocabularies, data models. Helps data producers share data so pay more attention to data curation and use elements of this framework.
- Long Tail Data
  - With the funding agency requirements on Data Management Plans and on making their data available, researchers more aware of importance of sharing data.
  - In astronomy, most original data from observations is in observatory archives, but at CDS we are seeing more data "attached to publications".



**Press Release 10-077**

# Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans

**Government-wide emphasis on community access to data supports substantive push toward more open sharing of research data**

**May 10, 2010**

During the May 5th meeting of the National Science Board, National Science Foundation (NSF) officials announced a change in the implementation of the existing policy on sharing research data. In particular, on or around October, 2010, NSF is planning to require that all proposals include a data management plan in the form of a two-page supplementary document. The research community will be informed of the specifics of the anticipated changes and the agency's expectations for the data management plans.

The changes are designed to address trends and needs in the modern era of data-driven science.

"Science is becoming data-intensive and collaborative," noted Ed Seidel, acting assistant director for NSF's Mathematical and Physical Sciences directorate. "Researchers from

# Progress in Environmental Data Curation?

Professor James Frew (UCSB):

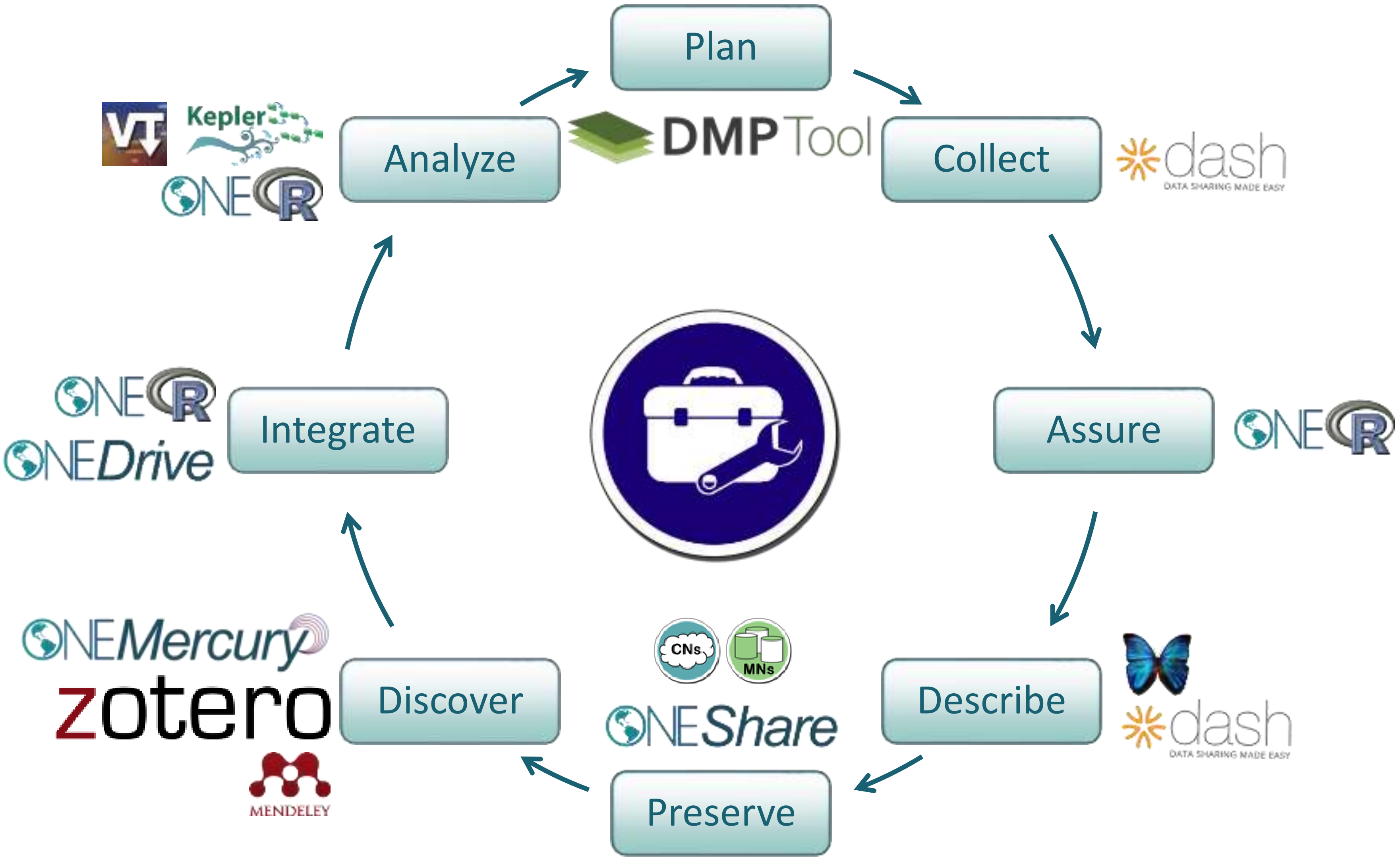
- Biggest change is funding agency mandate.
- NSF’s Data Management Plan for all proposals has made scientists (pretend?) to take data curation seriously.
- There are better curated databases and metadata now - but not sure that quality fraction is increasing!
- Frew’s first law: scientists don’t write 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

DataONE
Data Observation Network for Earth

# Enabling Science through Tools and Services



Plan

Collect

Assure

Describe

Preserve

Discover

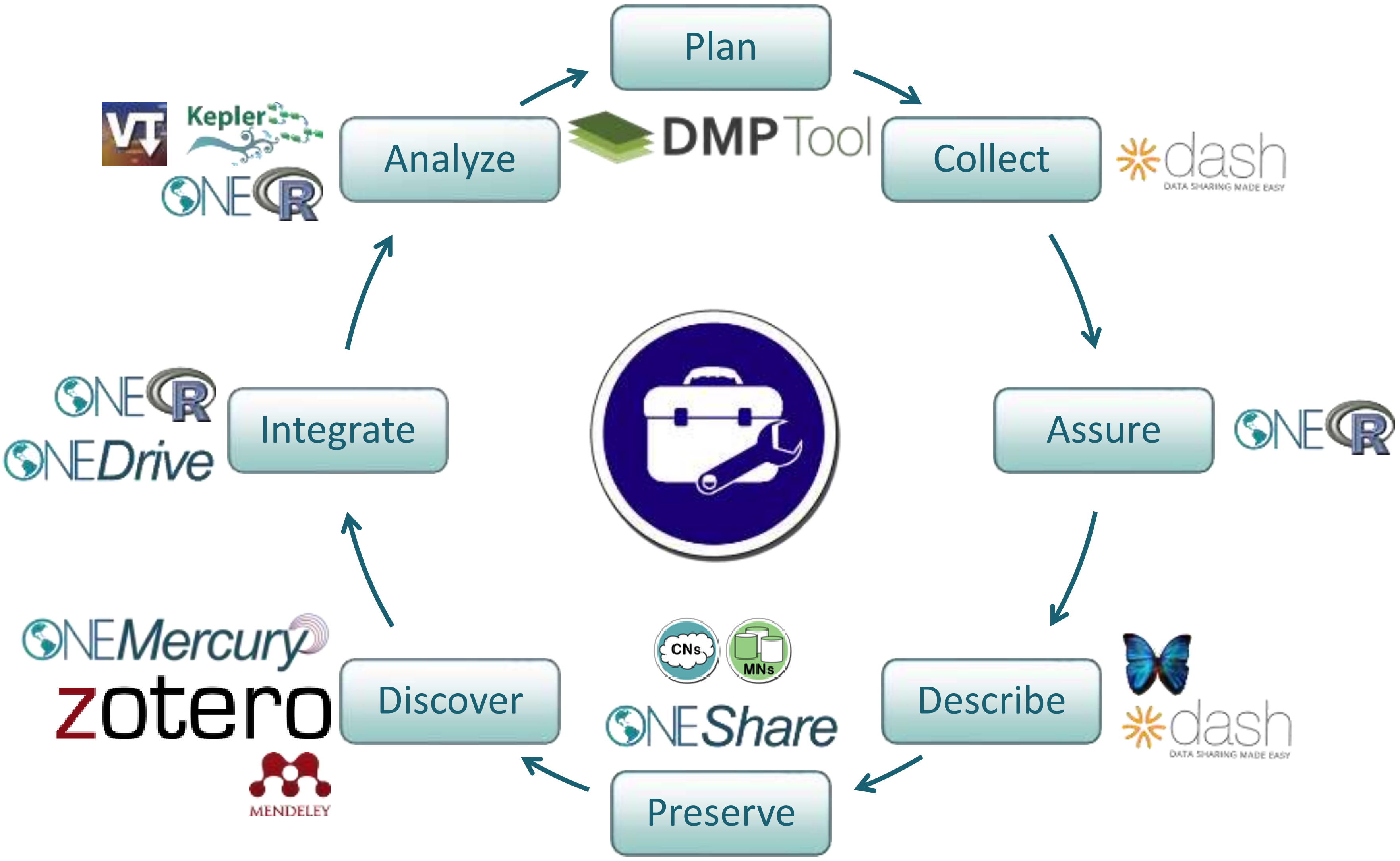
Integrate

Analyze

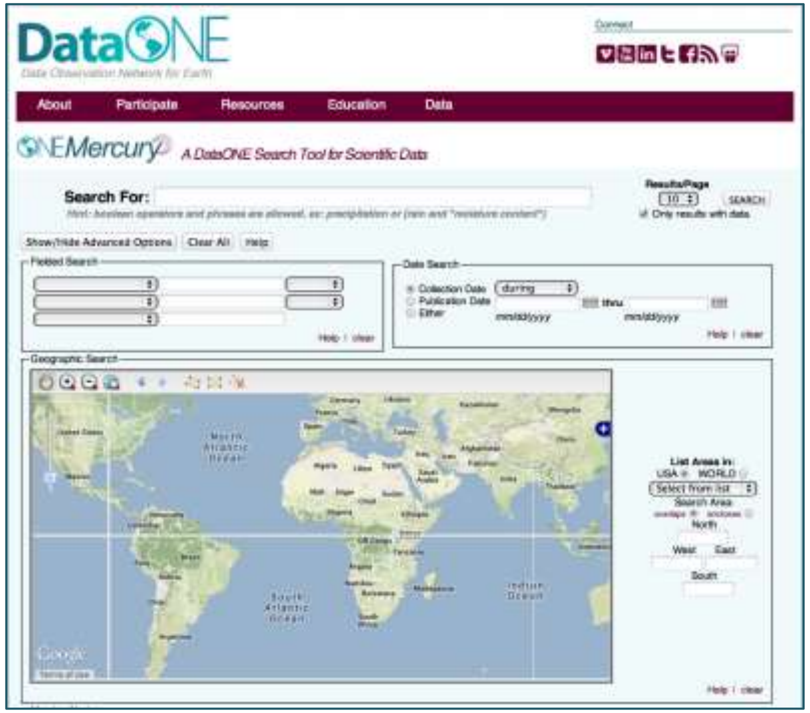
DataONE

Data Observation Network for Earth

# Enabling Science through Tools and Services



# DataONE: Data Management, Sharing and Publication


DataONE
Data Observation Network for Earth
Connect
About
Participate
Resources
Education
Data
ONEMercury
A DataONE Search Tool for Scientific Data

Search For:

Hint: boolean operators and phrases are allowed, ex: precipitation or (rain and "moisture content")

Results/Page 10 SEARCH

☑ Only results with data

Show/Hide Advanced Options | Clear All | Help

Fielded Search

Help | clear

Date Search

● Collection Date during

○ Publication Date mm/dd/yyyy thru mm/dd/yyyy

○ Either

Help | clear

Geographic Search

List Areas in: USA ○ WORLD ○

Select from list

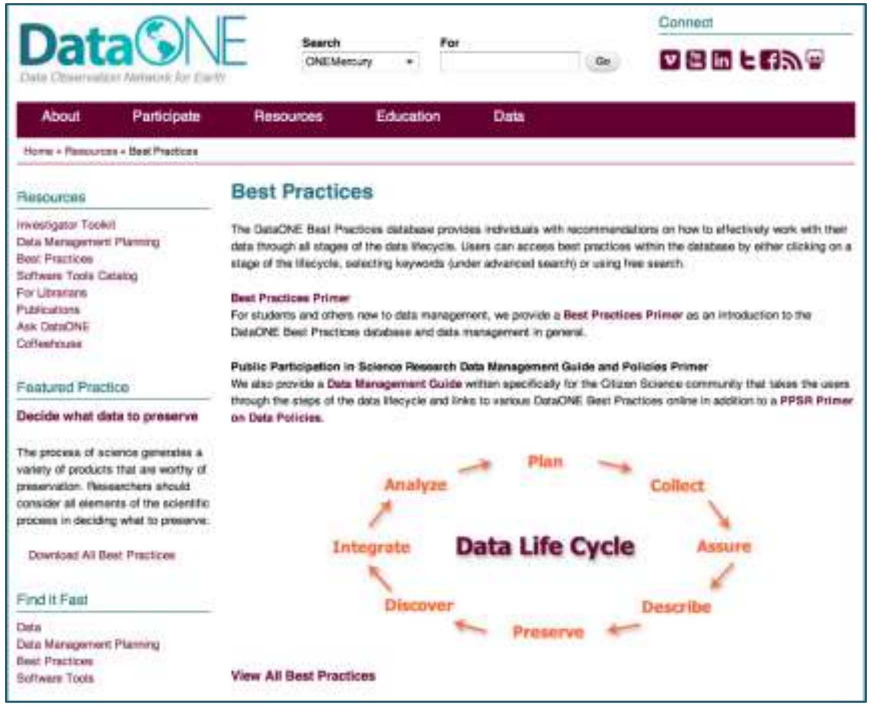
Search Area

North

West East

South

Help | clear



Home » Resources » Best Practices

Resources

- Investigator Toolkit
- Data Management Planning
- Best Practices
- Software Tools Catalog
- For Librarians
- Publications
- Ask DataONE
- Coffeehouse

Featured Practice

**Decide what data to preserve**

The process of science generates a variety of products that are worthy of preservation. Researchers should consider all elements of the scientific process in deciding what to preserve.

Download All Best Practices

Find It Fast

- Data
- Data Management Planning
- Best Practices
- Software Tools

## Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle. Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

**Best Practices Primer**

For students and others new to data management, we provide a **Best Practices Primer** as an introduction to the DataONE Best Practices database and data management in general.

**Public Participation in Science Research Data Management Guide and Policies Primer**

We also provide a **Data Management Guide** written specifically for the Citizen Science community that takes the users through the steps of the data lifecycle and links to various DataONE Best Practices online in addition to a **PPSR Primer on Data Policies.**

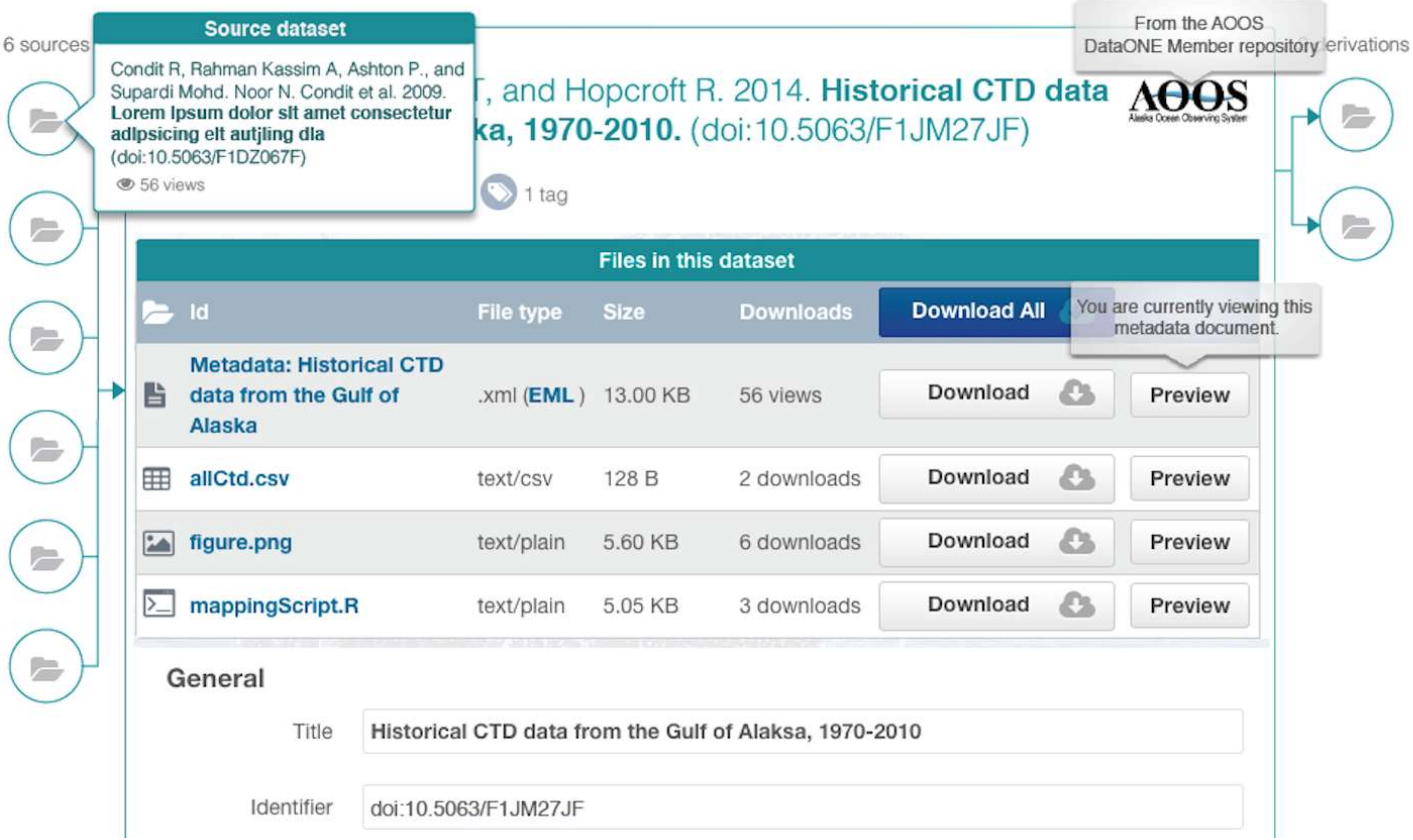
**View All Best Practices**

UTK and ORNL are partners in NSF DataONE Project

Slides courtesy of Bill Michener

Data Observation Network for Earth

# Provenance tracking and display



6 sources

**Source dataset**

Condit R, Rahman Kassim A, Ashton P., and Supardi Mohd. Noor N. Condit et al. 2009. **Lorem Ipsum dolor sit amet consectetur adipsicing elt autjling dla** (doi:10.5063/F1DZ067F)

56 views

From the AOOS DataONE Member repository

erivations

T, and Hopcroft R. 2014. **Historical CTD data** **ka, 1970-2010.** (doi:10.5063/F1JM27JF)

1 tag

**Files in this dataset**

| Id | File type | Size | Downloads | Download All | |
|---|---|---|---|---|---|
| Metadata: Historical CTD data from the Gulf of Alaska | .xml (EML ) | 13.00 KB | 56 views | Download | Preview |
| allCtd.csv | text/csv | 128 B | 2 downloads | Download | Preview |
| figure.png | text/plain | 5.60 KB | 6 downloads | Download | Preview |
| mappingScript.R | text/plain | 5.05 KB | 3 downloads | Download | Preview |

You are currently viewing this metadata document.

## General

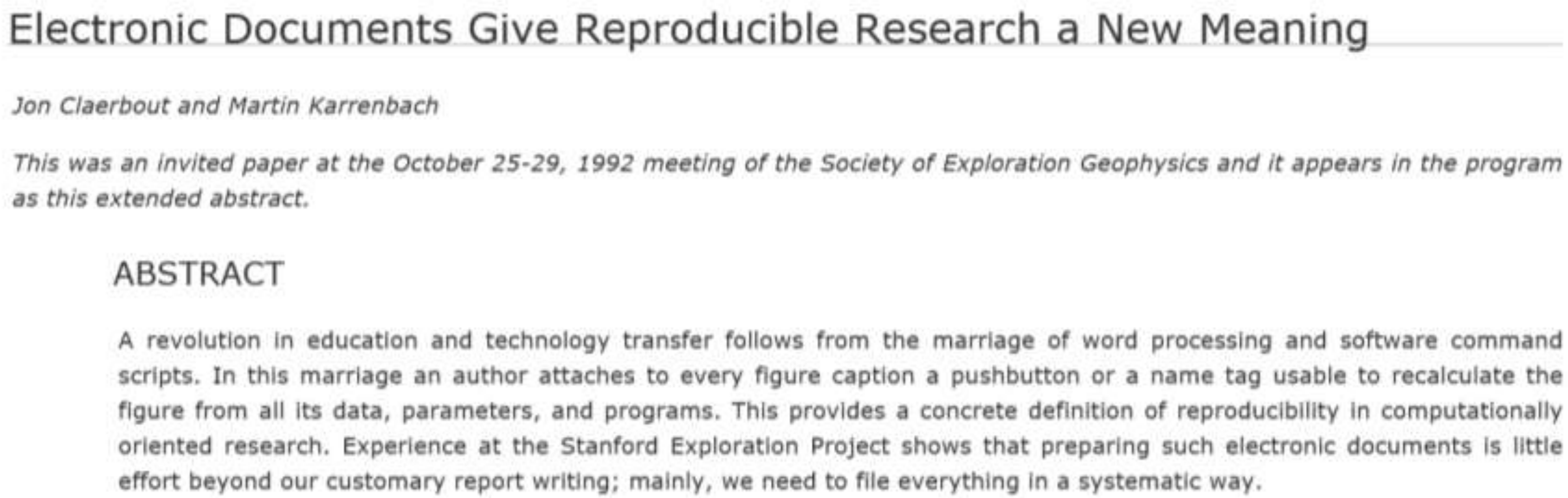
Title: Historical CTD data from the Gulf of Alaksa, 1970-2010

Identifier: doi:10.5063/F1JM27JF

# 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.

# 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

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. ■

# 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**

# Open Science Decoded

Nature Physics May 2015

See http://rdcu.be/cM1W

## Two  Technologies:

- □Machine Learning
- □Cloud Services

# Machine Learning

# Machine Learning

computers are great **tools** for

| | |
|---|---|
| storing | computing |
| managing | indexing |

huge amounts of **data**

we would like computers to also help with the **automatic**

| | |
|---|---|
| acquisition | discovery |
| aggregation | organization |
| correlation | analysis |
| interpretation | inference |

of the world's **information** and **knowledge**

# Machine Learning or ML

ML is exciting the IT industry since it enables us to:

- Build computing systems that improve with experience
- Solve extremely hard problems
- Extract more value from Big Data
- Approach human intelligence

In Science, Bayesian ML techniques help us manage uncertainties in data and assign probabilities to model predictions

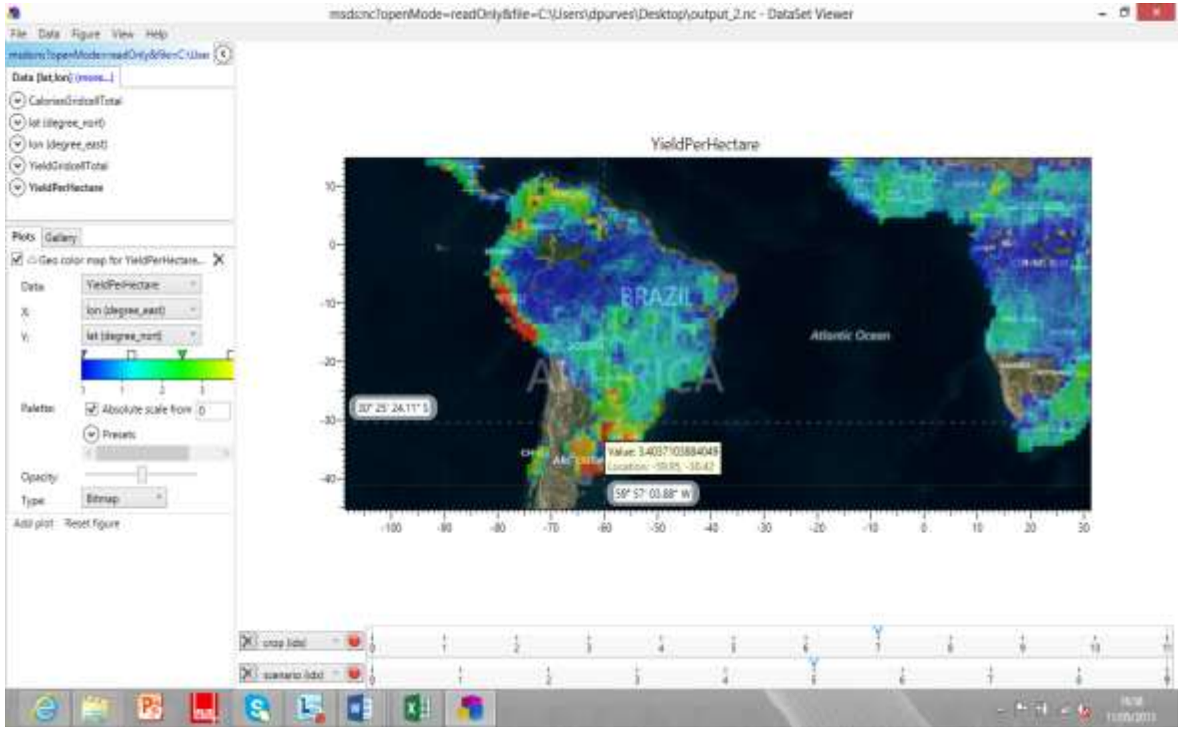
# Computational Ecology and Environmental Science

## Where environmental Questions meet Computer Science

Climate change: arguably greatest global challenge of the 21st century

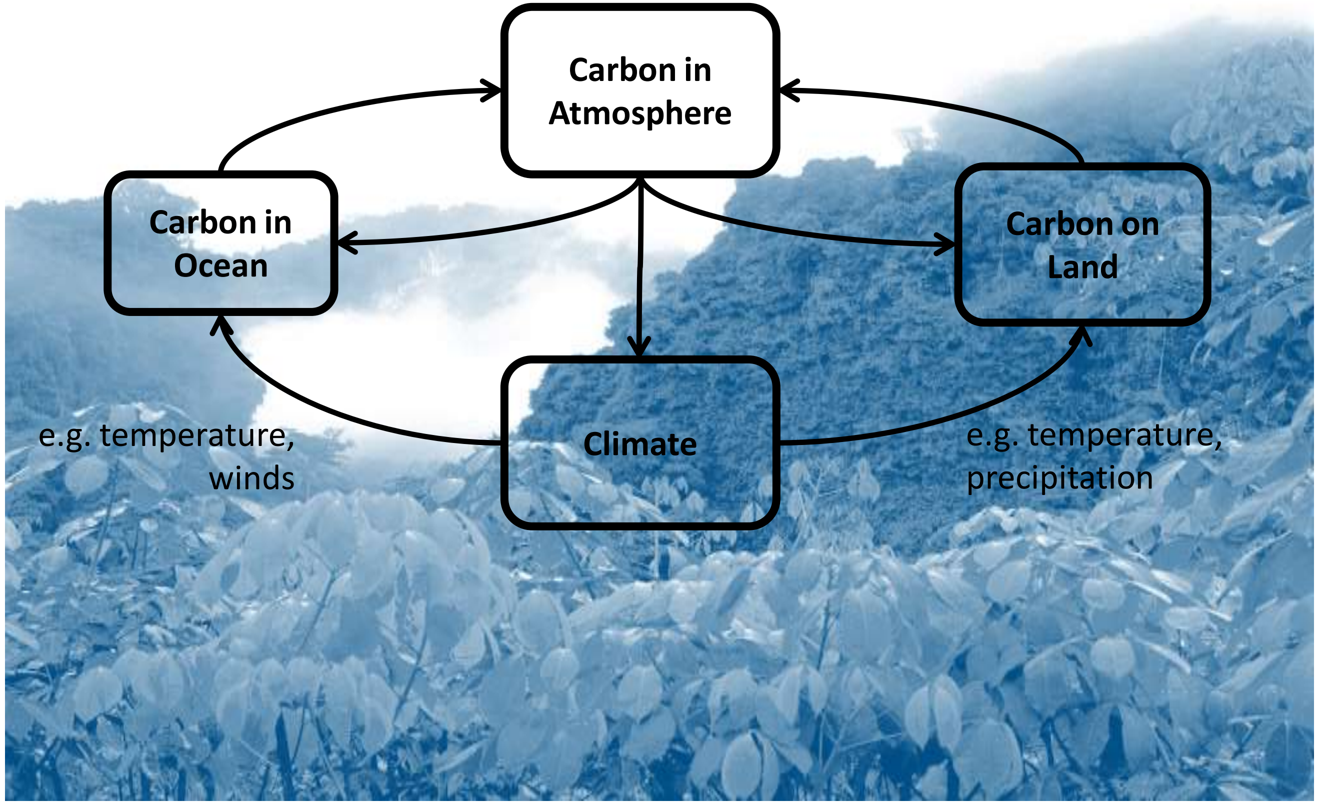
- What will be the impact of policy decisions and human actions?
- Will vegetation (mostly forests) continue to absorb 25% of human $CO_2$ emissions?

Computational Modeling seeks to create understanding from data sets that are distributed and diverse in time and space

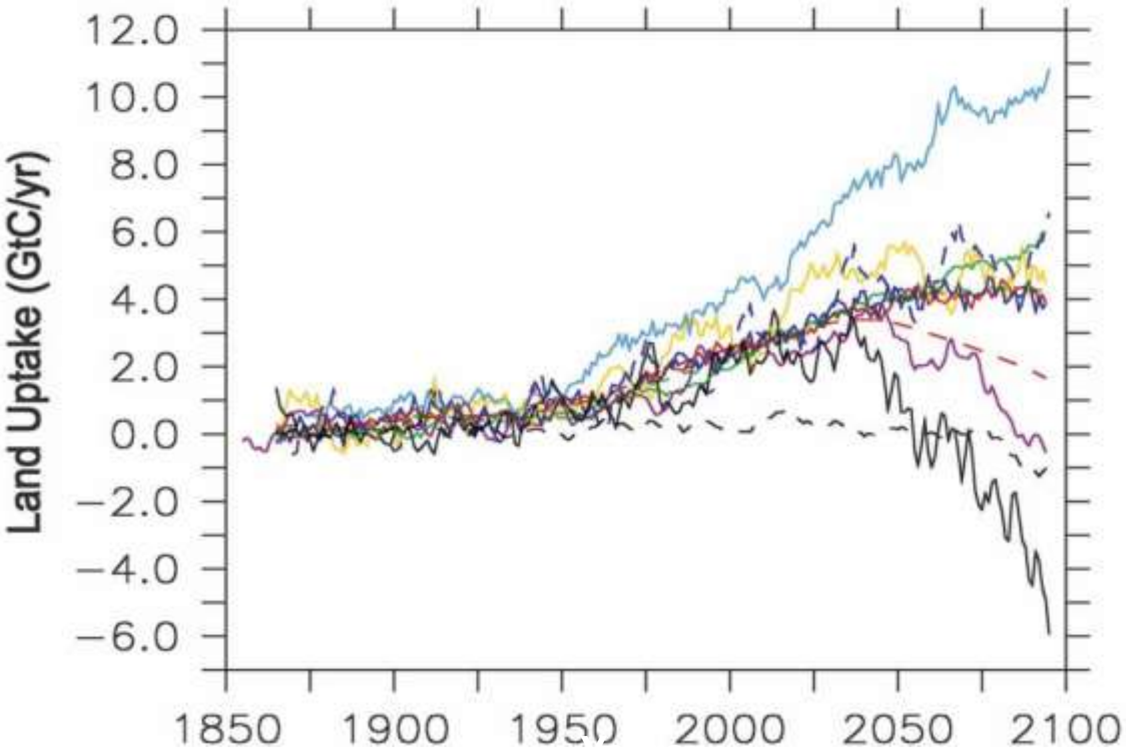


- *Drew Purves CEES Research Group*
  *Microsoft Research Cambridge, UK*

# Global Carbon-Climate Feedback Model



# Current Terrestrial Carbon Models are limited



# Dealing with Uncertainty

- Characterize and evaluate the component models of the Carbon-Climate feedback cycle
- Use Bayesian inference to characterize missing data and uncertainty in predictions

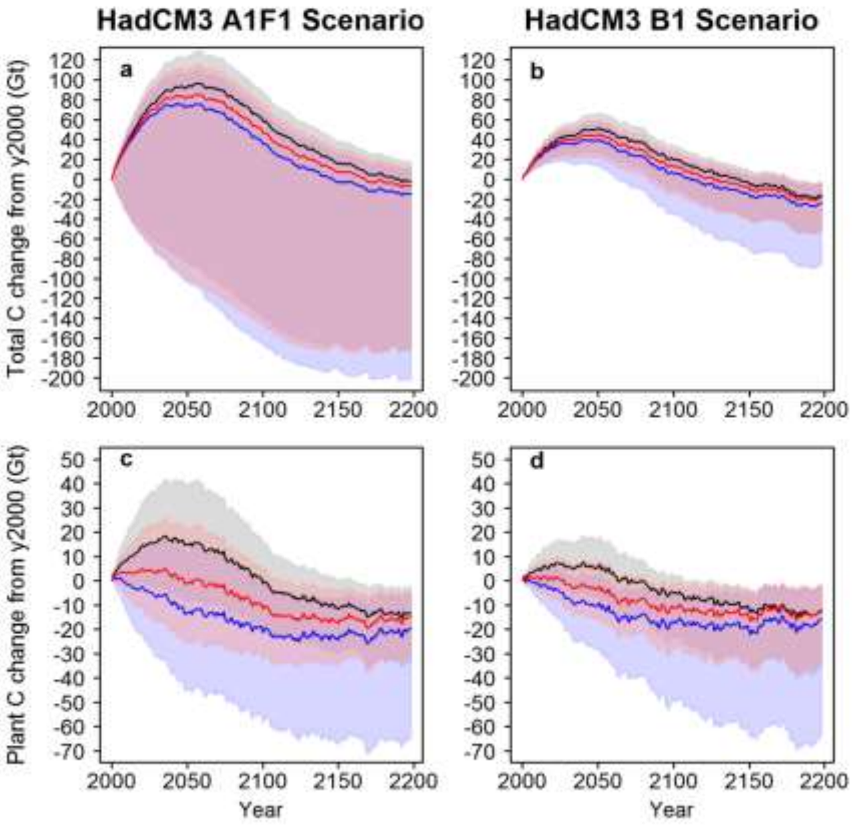
➢ Constrain parameters of component eco-physiological processes of a Dynamic Global Vegetation Model (DGVM) using 12 global empirical data sets

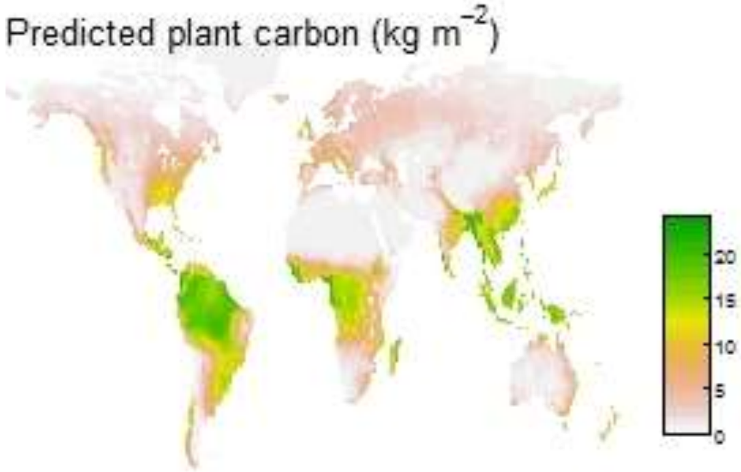
➢ See paper by Matthew Smith et al. :

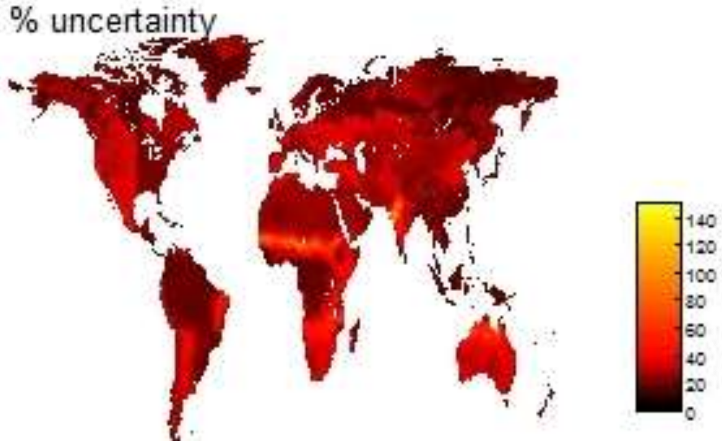
**“The climate dependence of the terrestrial carbon cycle; including parameter and structural uncertainties”**

**http://www.biogeosciences-discuss.net/9/13439/2012/bgd-9-13439-2012.html**

# Future Global Changes Predicted with Uncertainty Ranges

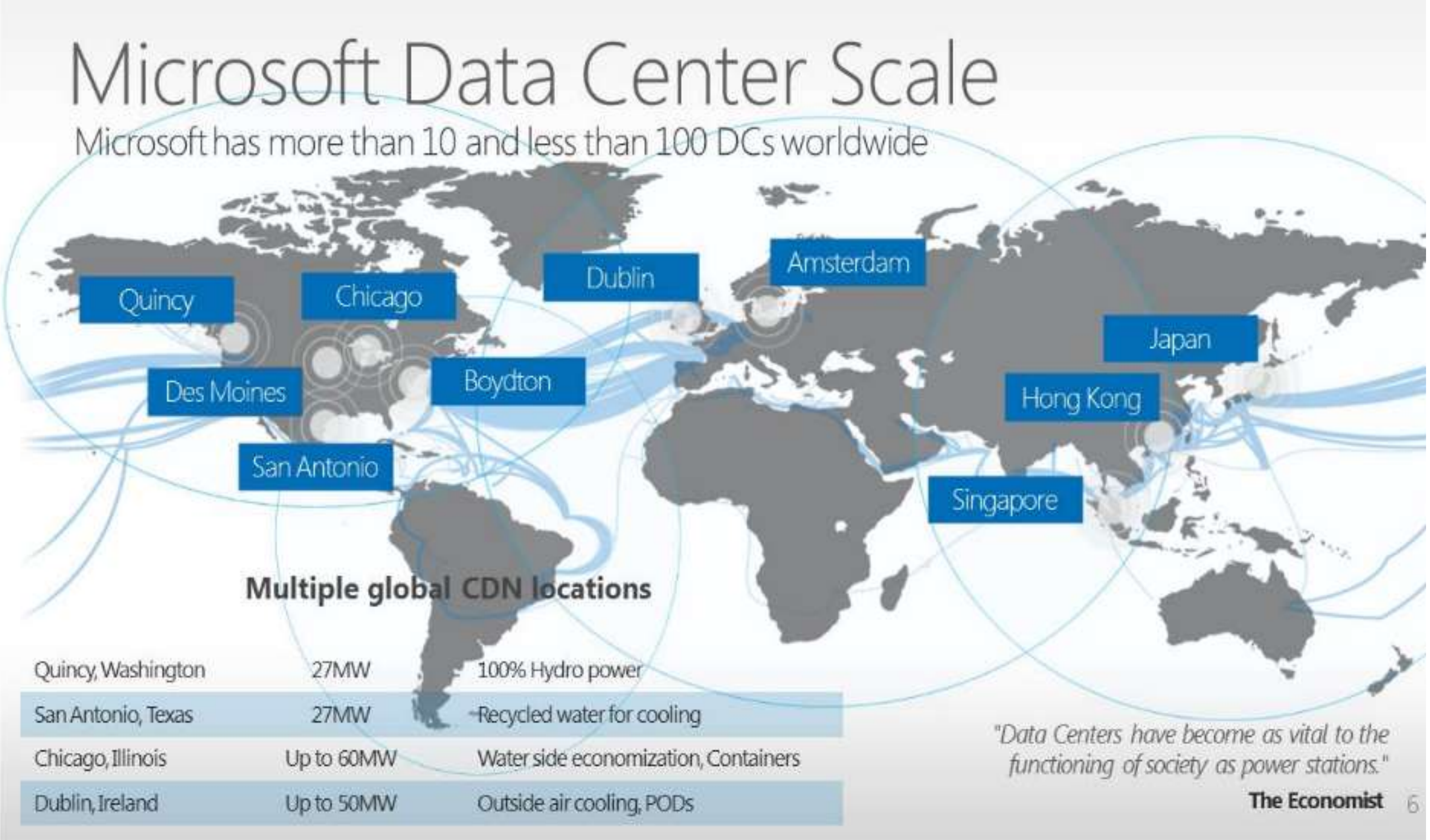






# Cloud Services

# Industry is building out massive Infrastructure



## Microsoft Data Center Scale

Microsoft has more than 10 and less than 100 DCs worldwide

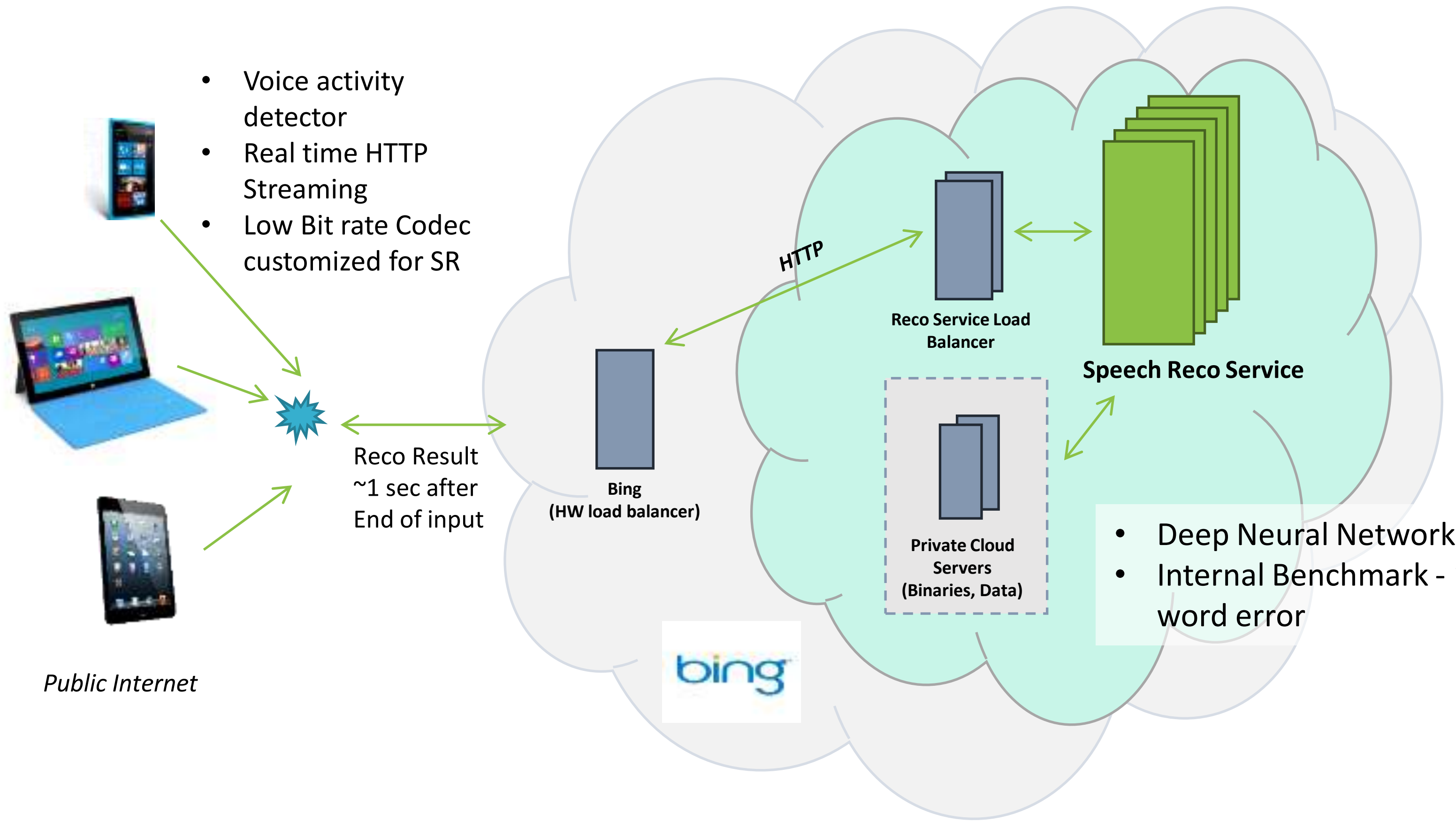
**Multiple global CDN locations**

| | | |
|---|---|---|
| Quincy, Washington | 27MW | 100% Hydro power |
| San Antonio, Texas | 27MW | Recycled water for cooling |
| Chicago, Illinois | Up to 60MW | Water side economization, Containers |
| Dublin, Ireland | Up to 50MW | Outside air cooling, PODs |

> *"Data Centers have become as vital to the functioning of society as power stations."*
>
> **The Economist**

# Bing Speech Recognition Service Architecture

- Voice activity detector
- Real time HTTP Streaming
- Low Bit rate Codec customized for SR



- Deep Neural Networks
- Internal Benchmark - 13.5% word error

# The Cloud Advantages

## Omnipresent Services

- Uploading data
- Download commands
- Streaming signals
- Network between Devices

## Compute and Storage Elasticity

- Lower barriers to adoption
- Lower barriers to scaling
- Lower overheads

## Accelerates Collaboration

- Sharing data
- Sharing algorithms
- Co-authoring
- Reproducible Research

## Cloud is not High-End Supercomputing

- Cloud offers modest HPC services
- Supports Hadoop, Apache Spark, …
- Machine Learning services for Big Data

# An Example of Cloud Power:
# Genomics and Personalized medicine

Use genetic markers to…

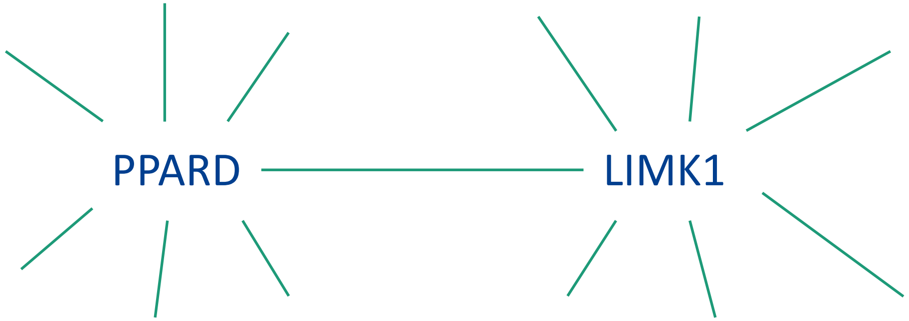
- Understand causes of disease
- Diagnose a disease
- Infer propensity to get a disease
- Predict reaction to a drug

# ‘Moondog’ Azure Cloud Project

- Wellcome Trust data for seven common diseases
- With FaST-LMM and Azure, can look at all SNP pairs (about 60 billion of them)
- 1,000 compute years; 20 TB output
- Using 27,000 Azure compute cores analysis was completed in just 13 days

# Moondog project with Azure: First Results

In coronary artery disease…

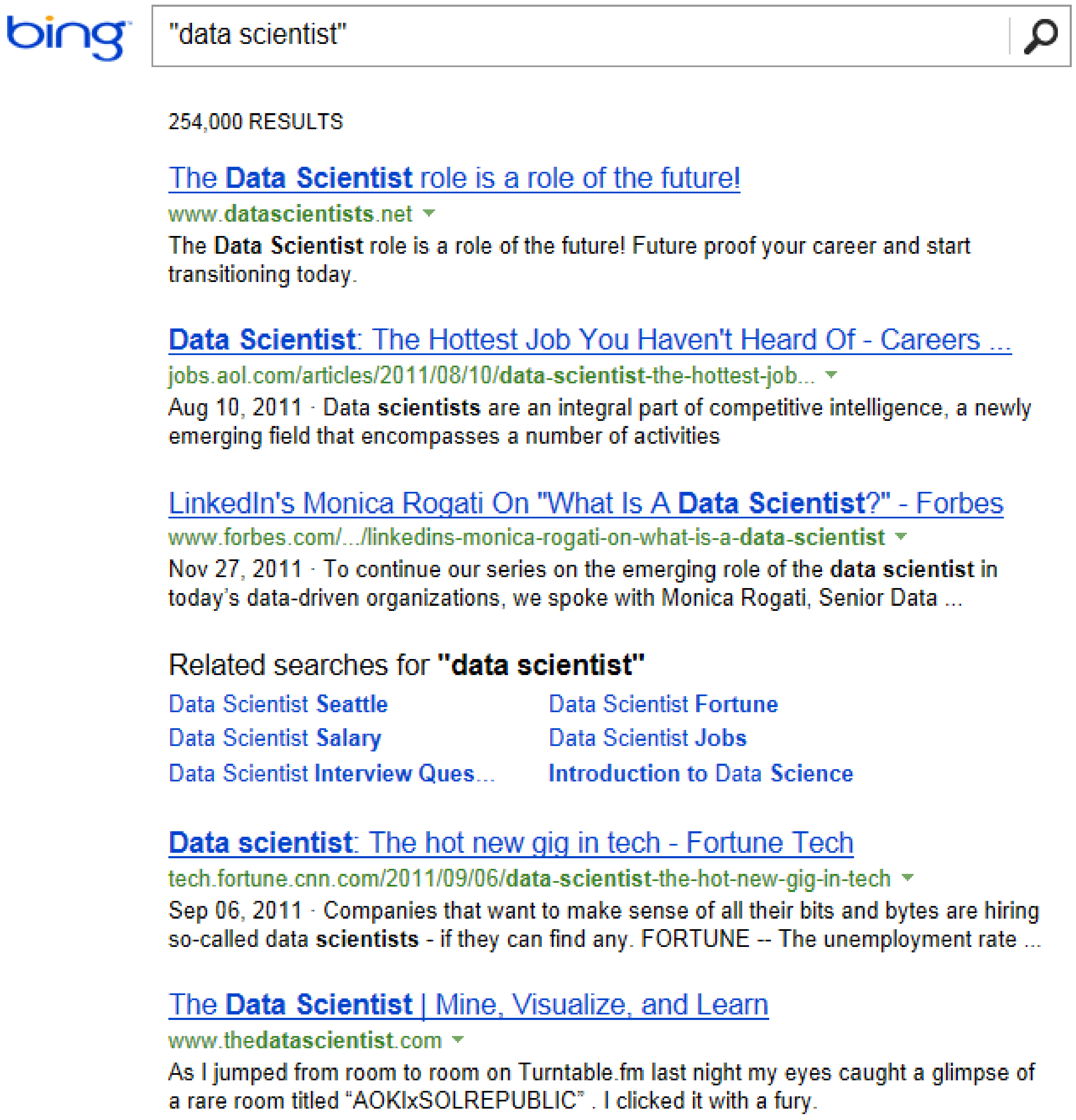


SCIENTIFIC REPORTS

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

# Data Science in the Future?



"data scientist"

254,000 RESULTS

**The Data Scientist role is a role of the future!**
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...
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
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** |

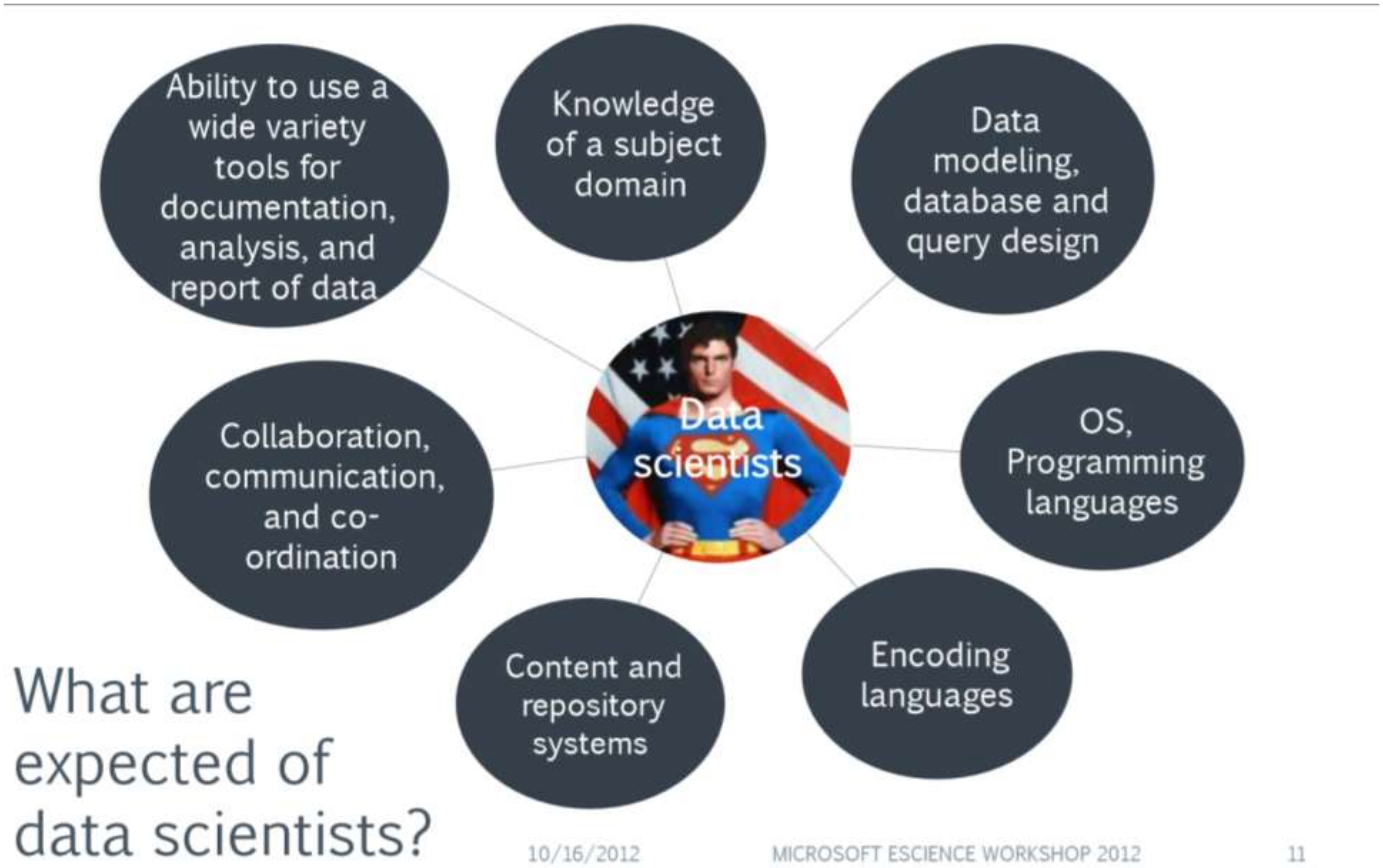
**Data scientist: The hot new gig in tech - Fortune Tech**
tech.fortune.cnn.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech
Sep 06, 2011 · Companies that want to make sense of all their bits and bytes are hiring so-called data **scientists** - if they can find any. FORTUNE -- The unemployment rate ...

**The Data Scientist | Mine, Visualize, and Learn**
www.thedatascientist.com
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**<br> | **People who are expert at**<br>• Operating at low levels close to the data, write code that manipulates<br>• They may have some machine learning background.<br>• Large companies may have teams of them in-house or they may look to third party specialists to do the work. |
| **Data Analyst**<br> | **People who explore data through statistical and analytical methods**<br>• They may know programming;  May be an spreadsheet wizard.<br>• Either way, they can build models based on low-level data.<br>• They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these. |
| **Data Steward**<br> | **People who think to managing, curating, and preserving data.**<br>• They are information specialists, archivists, librarians and compliance officers.<br>• 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*



Ability to use a wide variety tools for documentation, analysis, and report of data

Knowledge of a subject domain

Data modeling, database and query design

Collaboration, communication, and co-ordination

Data scientists

OS, Programming languages

Content and repository systems

Encoding languages

# What are expected of data scientists?

10/16/2012 MICROSOFT ESCIENCE WORKSHOP 2012 11

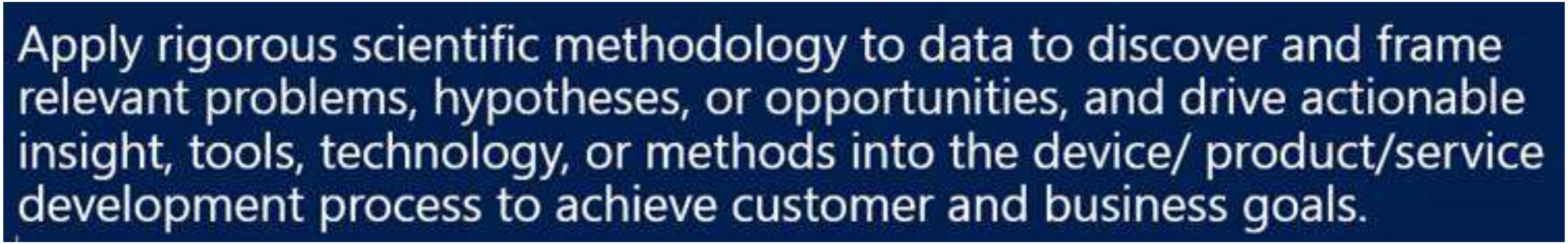
Slide courtesy of Jian Qin

# Microsoft – new roles for Data Scientists

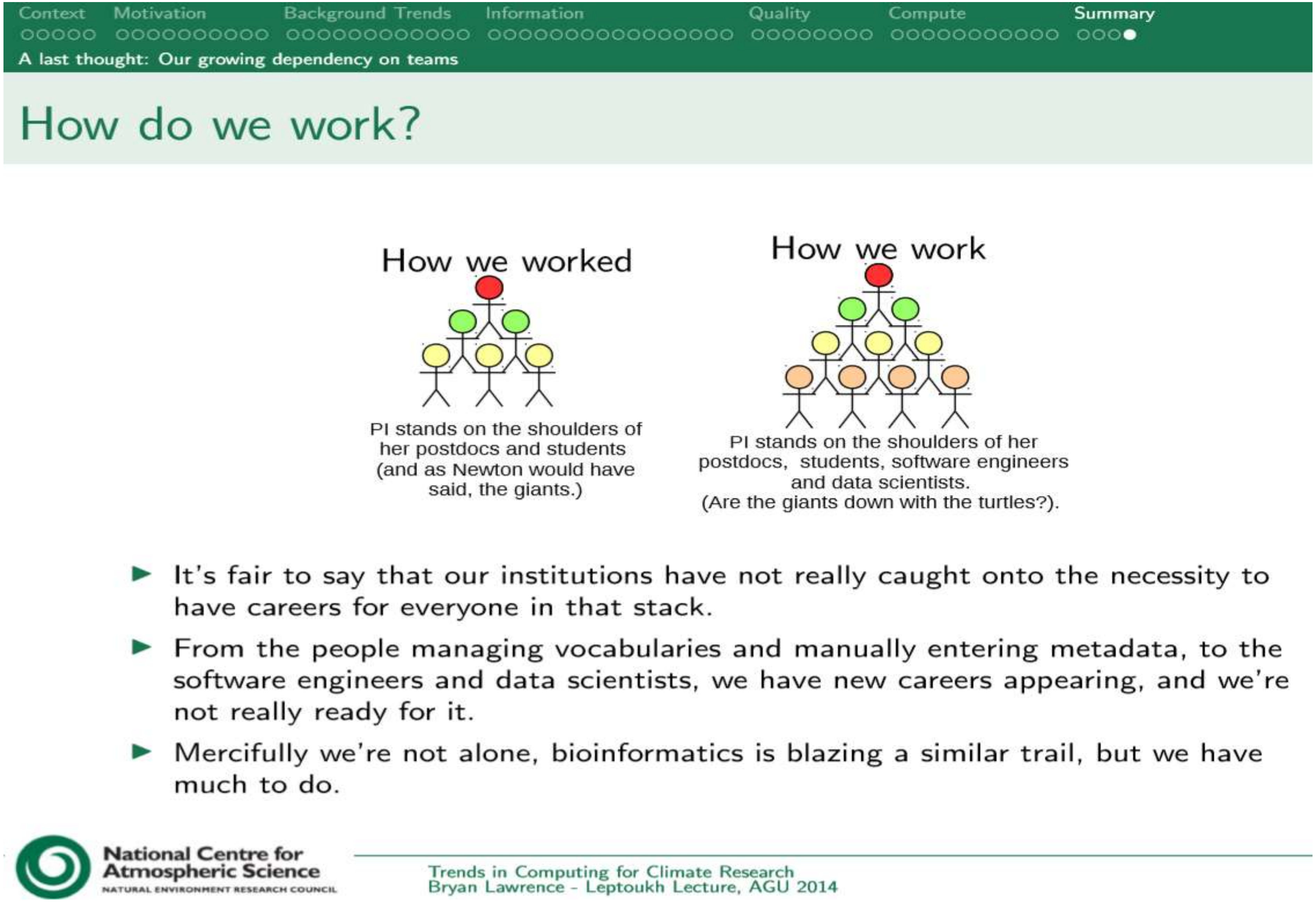
DATA & APPLIED SCIENTIST

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.



# 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

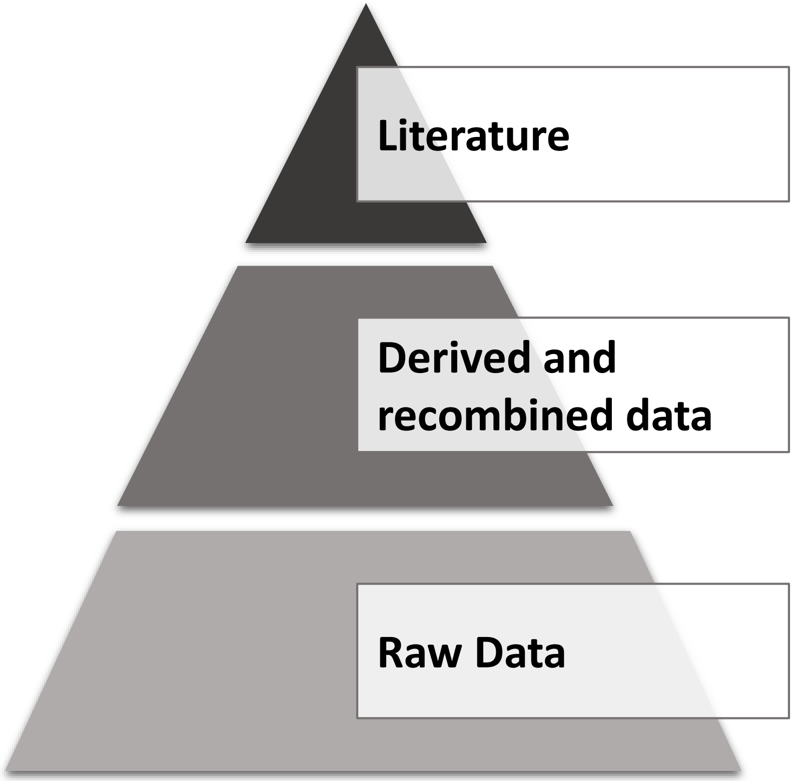
PI stands on the shoulders of her postdocs, students, software engineers and data scientists.
(Are the giants down with the turtles?).

- 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.

Slide thanks to Bryan Lawrence

# Jim Gray’s Vision: All Scientific Data Online

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



Literature

Derived and recombined data

Raw Data

*(From Jim Gray’s last talk)*