Open Science and e-Infrastructure

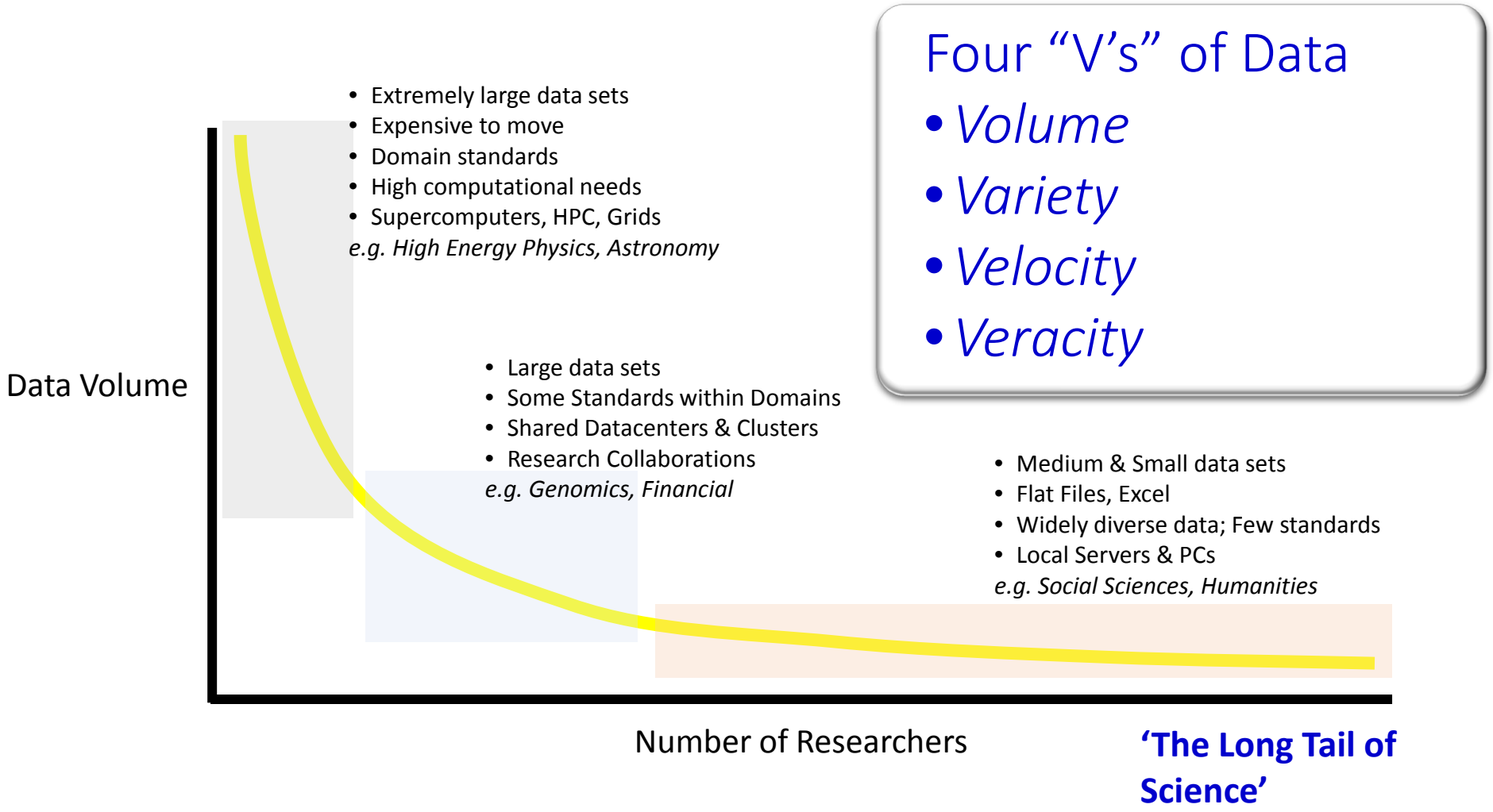
Professor Tony Hey
Chief Data Scientist
Science and Technology Facilities Council
Department of Business, Innovation and Skills, UK

Outline

- Fourth Paradigm: Data-intensive science
  - Astronomy, Genetics and Environmental Science
- Open Access, Open Data and Open Science
  - Budapest and Berlin declarations
  - White House Memo in the US
  - Reproducible Research
- Network requirements for Data-Intensive Science
  - TCP and end-to-end performance
  - Science DMZs and Superfacilities
  - Use by industry
- Industry, Data Scientists and e-Infrastructure
  - Training data scientists
  - Access to scientific infrastructure by industry

The Fourth Paradigm: Data-Intensive Science

Much of Science is now Data-Intensive



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 before analysis by astronomers

➢ Platform for ‘citizen science’ with GalaxyZoo project

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…

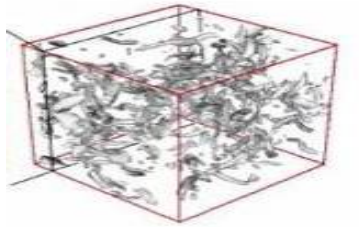
Last few decades – **Computational Science**

- 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

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



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

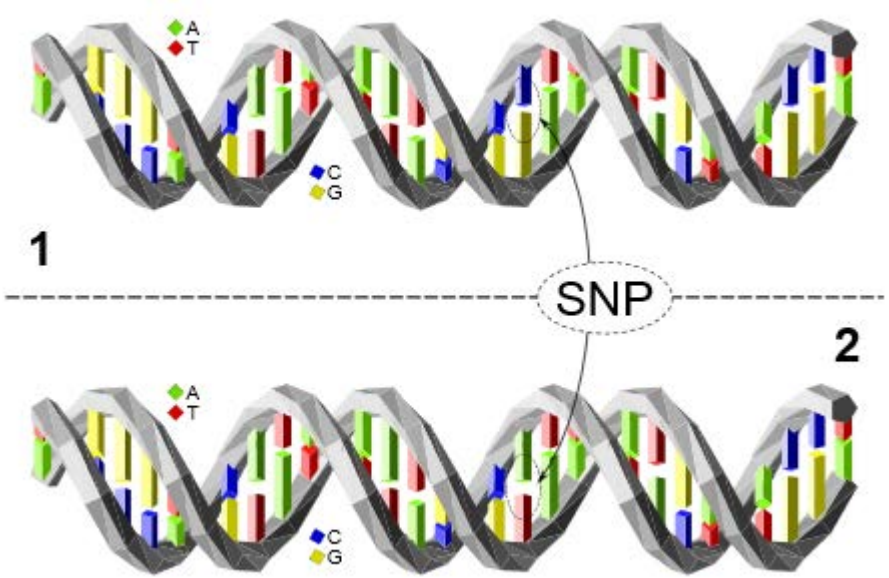
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

(With thanks to Jim Gray)

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

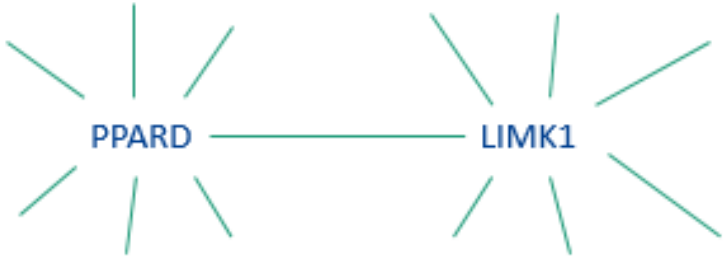


Genomics, Machine Learning and the Cloud

The Problem

- Wellcome Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls
- 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 data
- 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

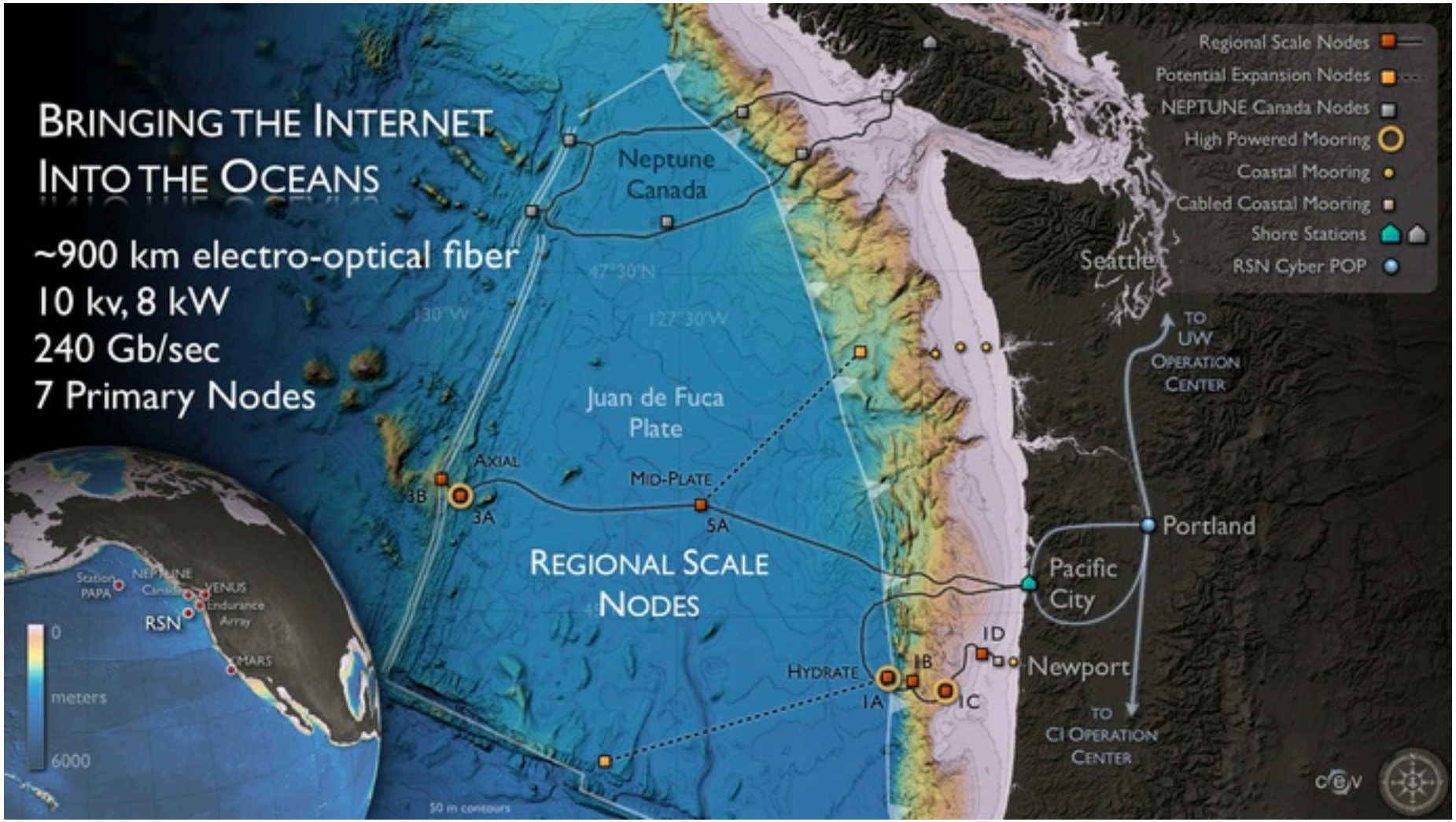


SCIENTIFIC REPORTS

An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data

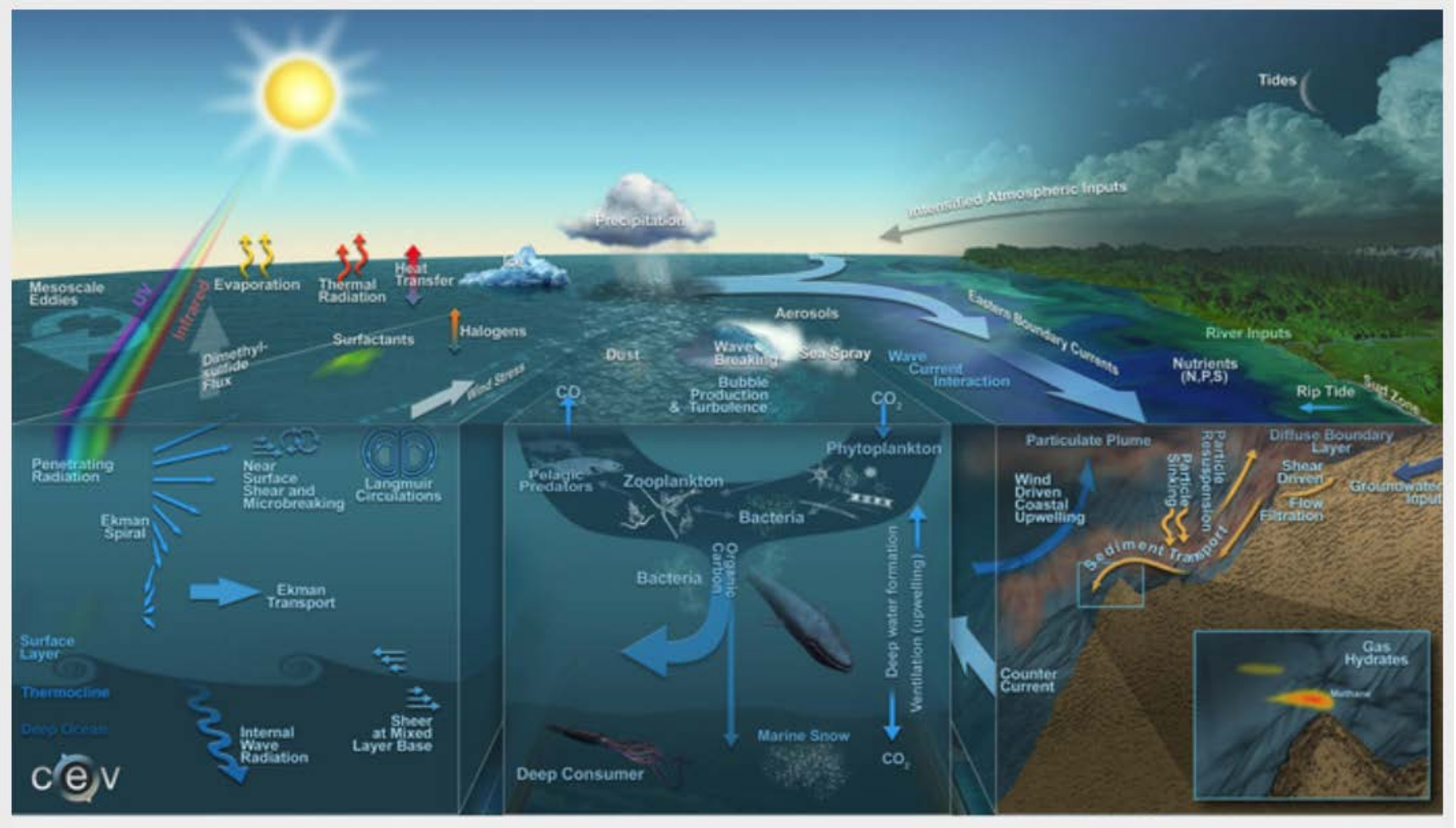
Christoph Lippert, Jennifer Listgarten, Robert I. Davidson, Jeff Baxter, Hoifung Poon, Carl M. Kadie & David Heckerman

NSF's Ocean Observatory Initiative



Slide courtesy of John Delaney

Oceans and Life



Slide courtesy of John Delaney

Open Access, Open Data and Open Science

The Budapest Open Access Initiative (2001)

- The Budapest Open Access Initiative came from a meeting convened in Budapest by the Soros's Open Society Institute in December 2001
- The purpose of the meeting was to accelerate the international effort to make research articles in all academic fields freely available on the Internet
- First to give a definition of 'Open Access'

The Berlin Declaration 2003

- ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’

- Defined open access contributions as including:

  ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’

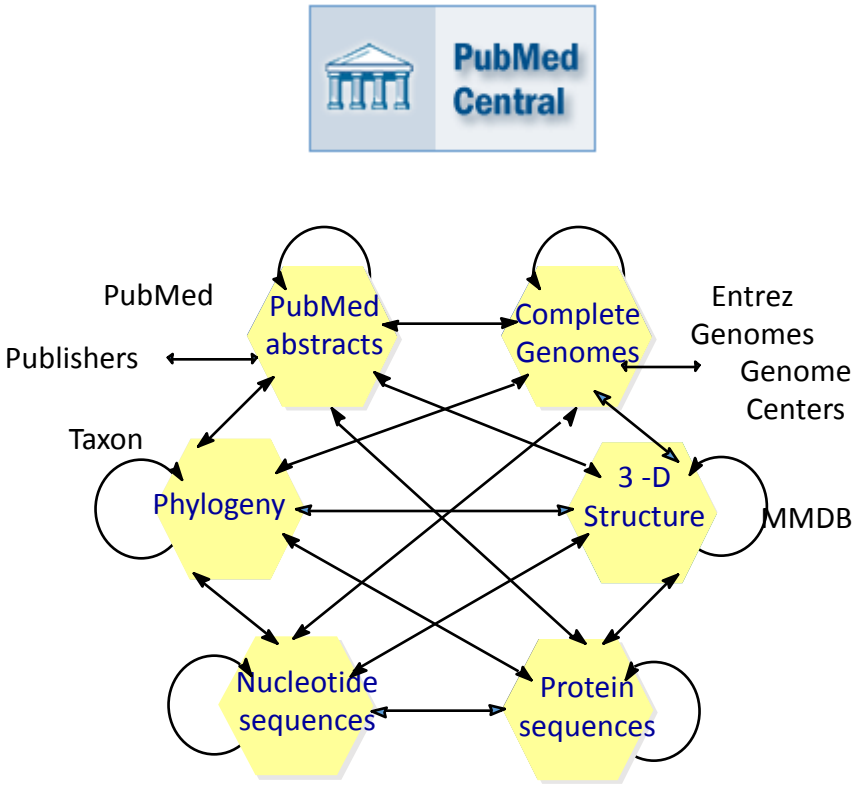
US White House Memo on increased public access to the results of federally-funded research

- Directive required 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 memo defines research results to encompass not only the research paper but also *"… 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 …."*

22 February 2013

The US National Library of Medicine

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

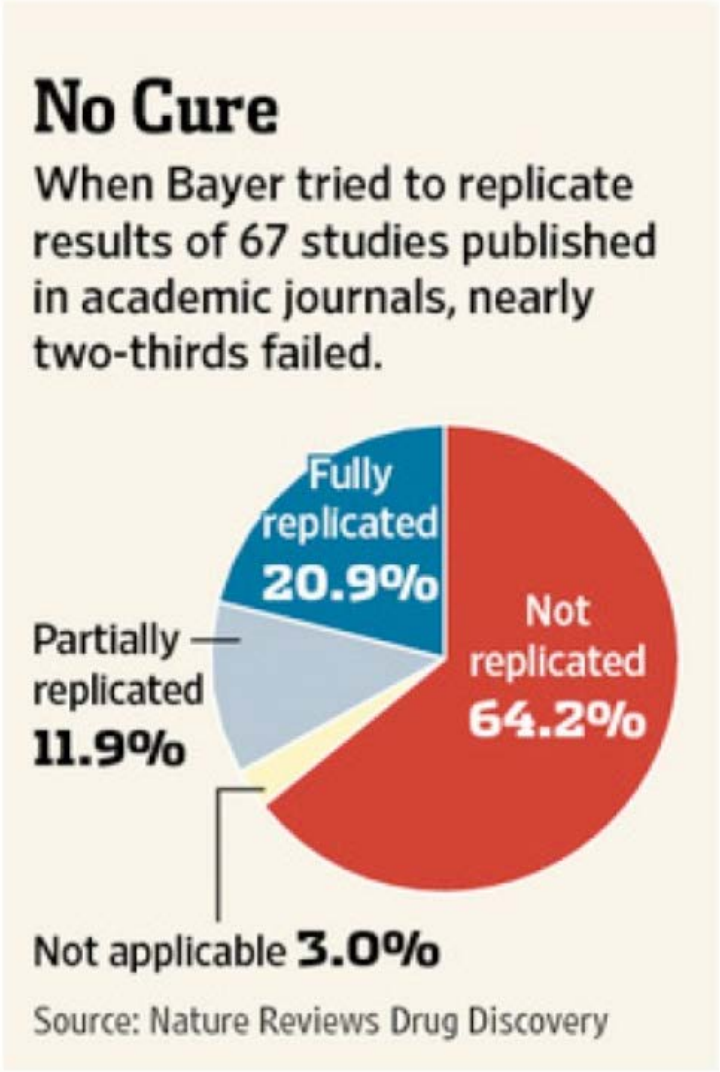


Entrez cross-database search

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!



Sustainability of Data Links?

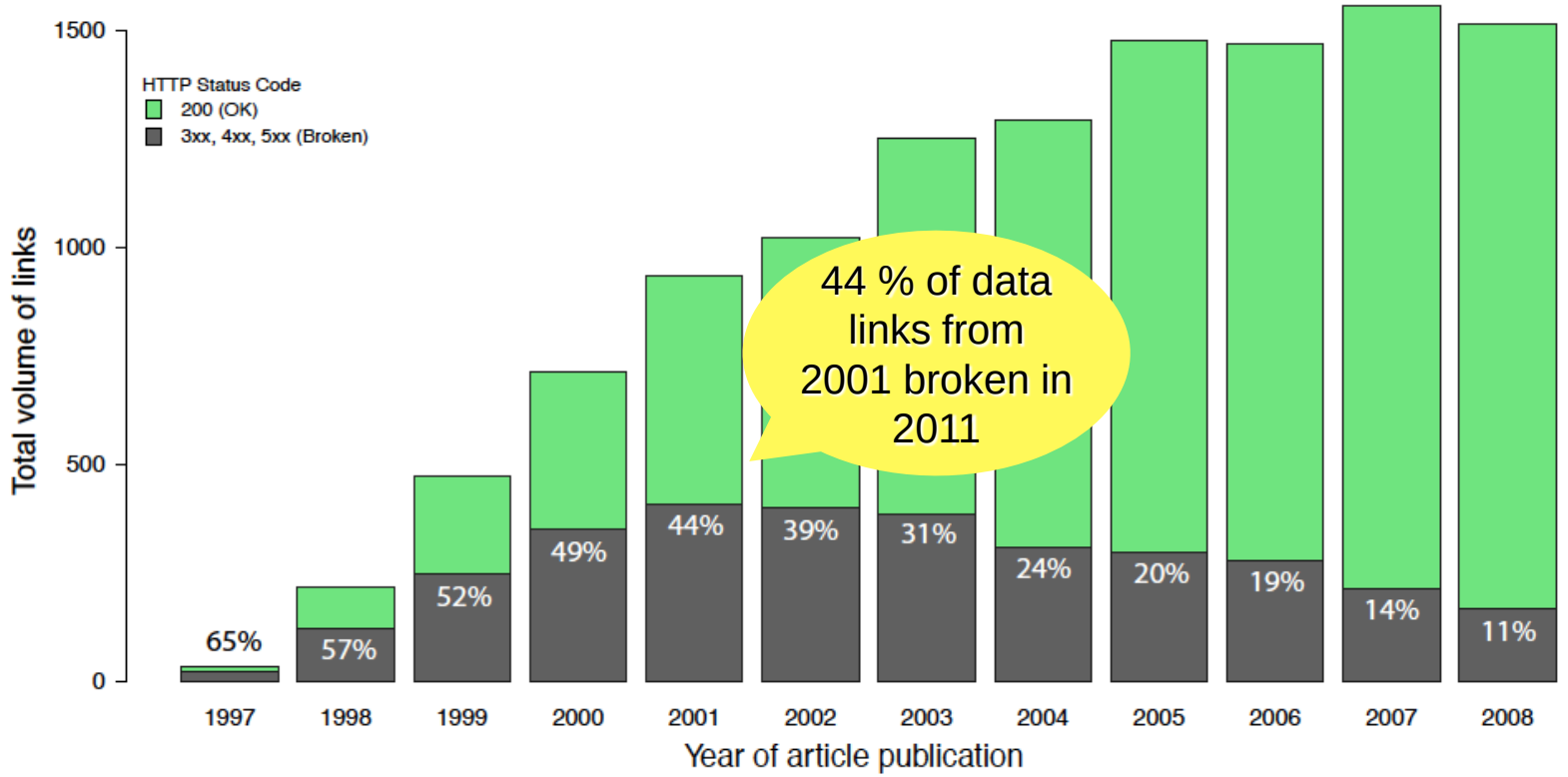


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

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.

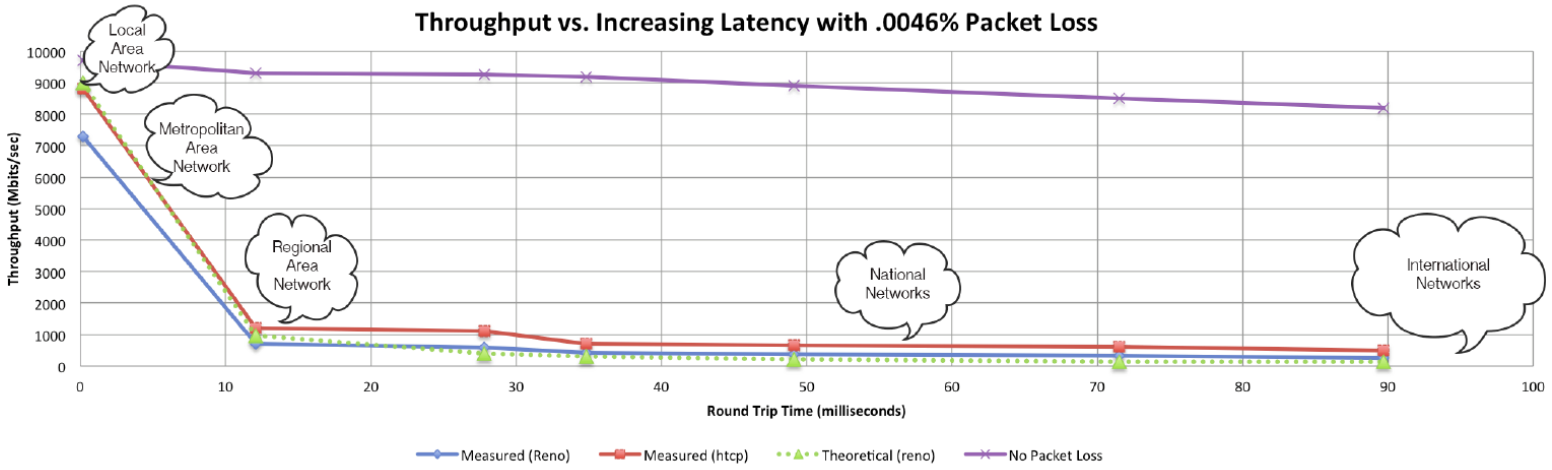
ORCID - Open Research & Contributor ID

- 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

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

The Problem …

- 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

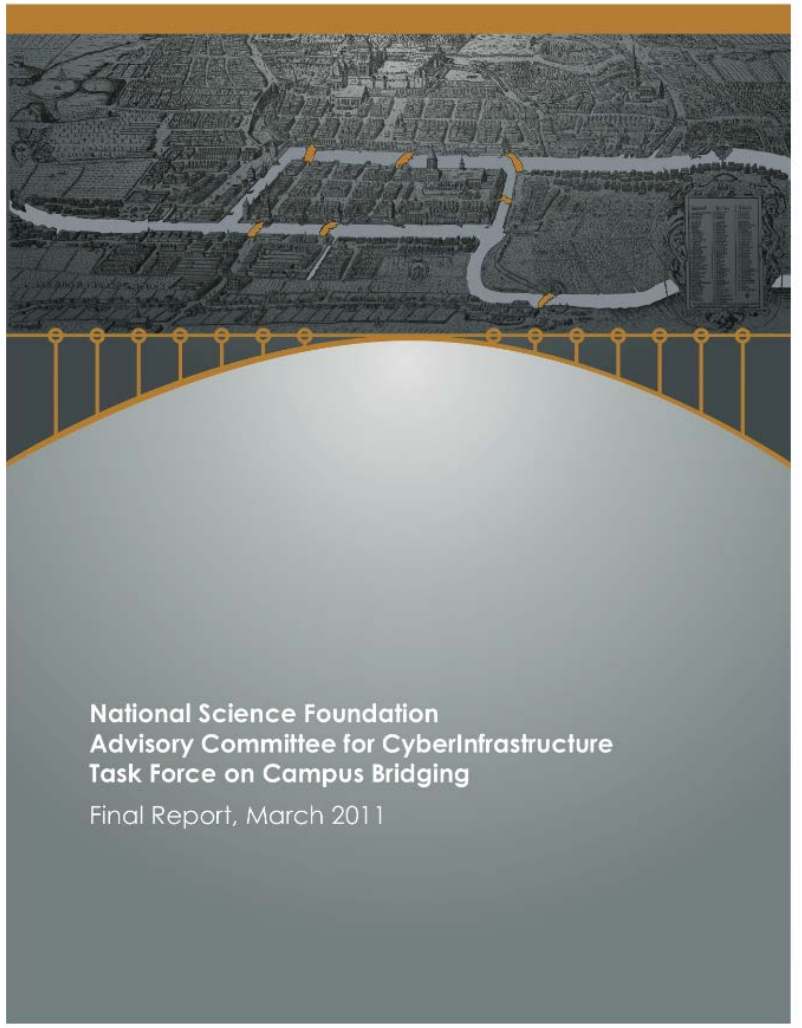
Thanks to Eli Dart, LBNL

NSF Task Force on 'Campus Bridging' (2011)

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

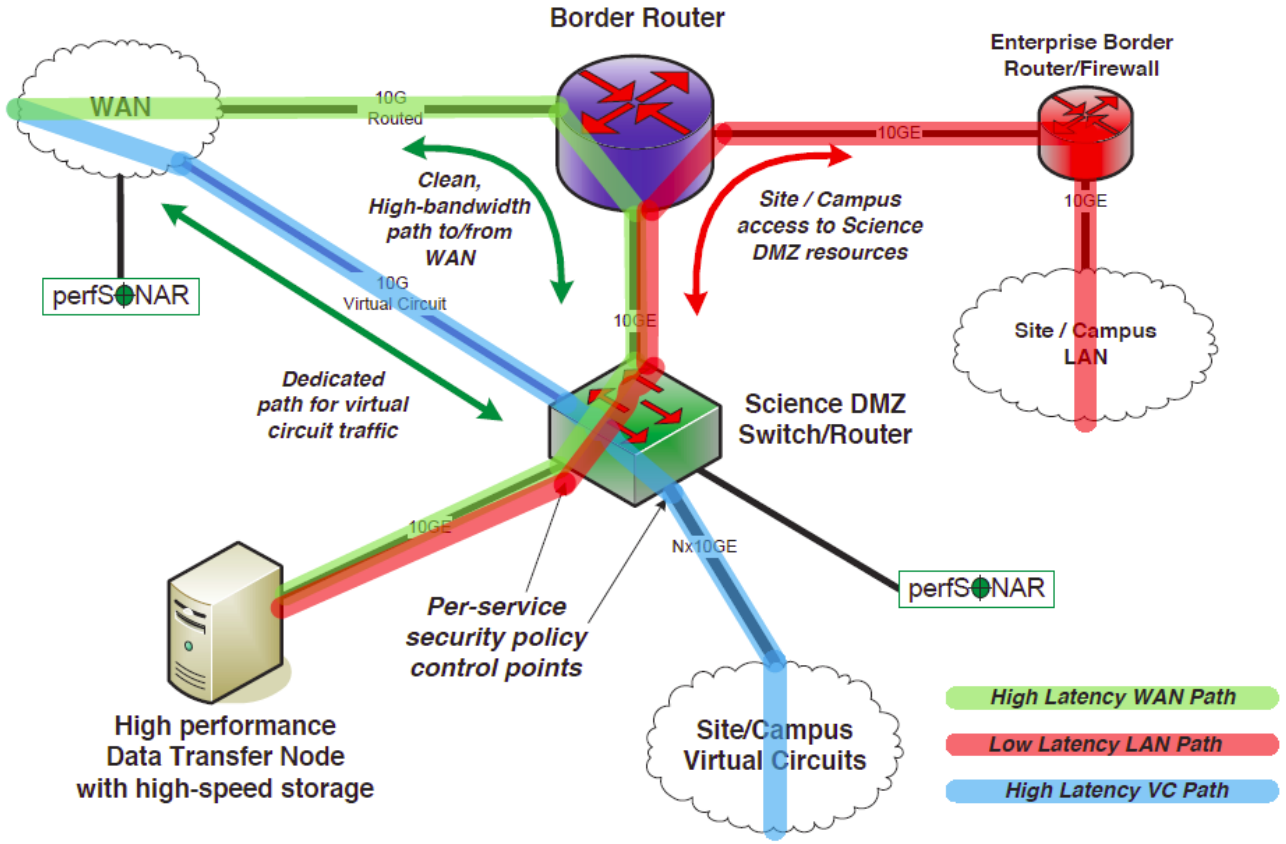
- a researcher's personal cyberinfrastructure
- cyberinfrastructure at other campuses
- cyberinfrastructure at the regional, national and international levels

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



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



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

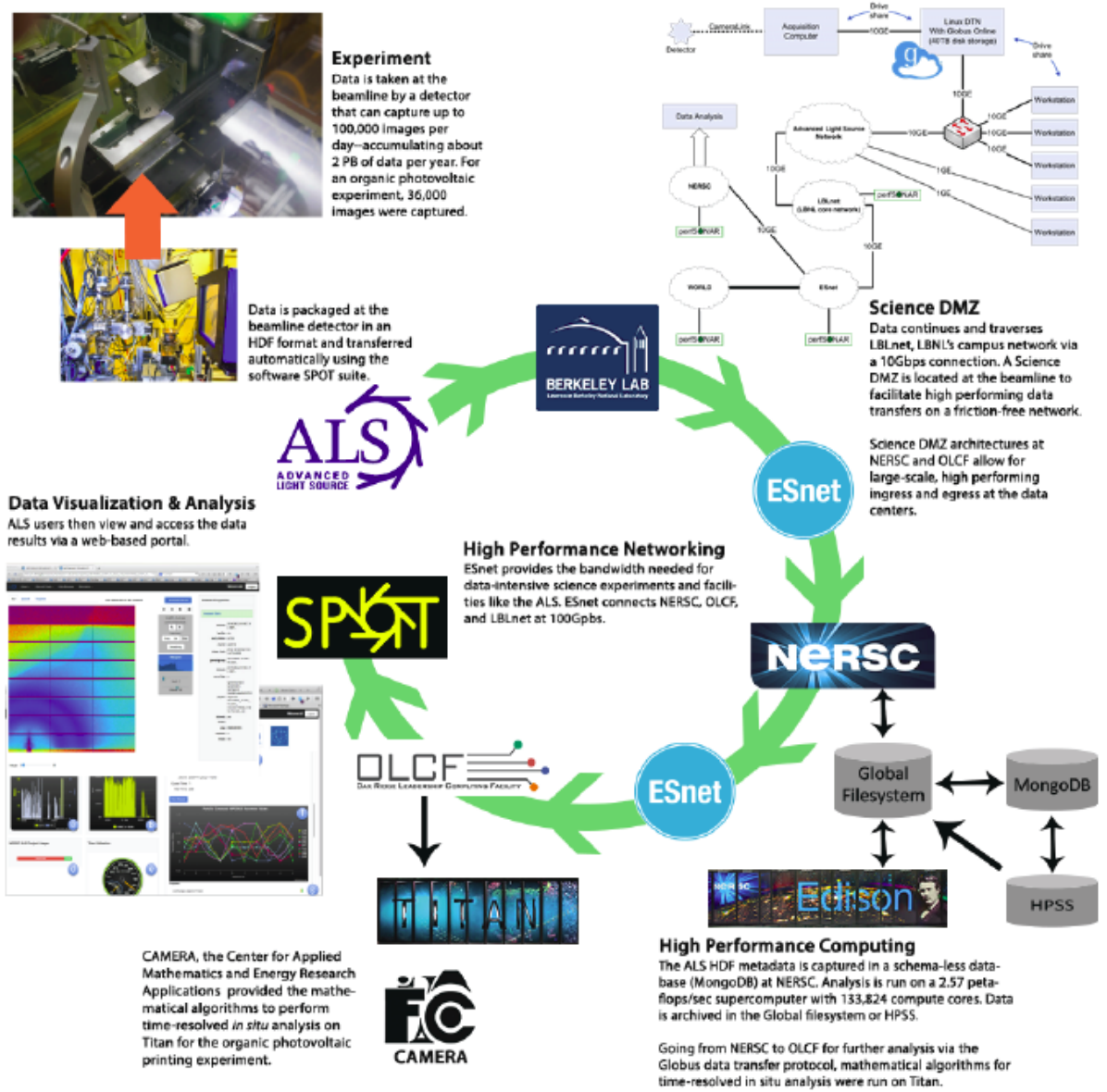
- ➢ Need to connect ESFRI Large Experimental Facilities and HPC systems via Science DMZs
- ➢ Need research funding agencies to work together with GEANT and NRENs to support high bandwidth end-to-end connections to researchers at institutions
- ➢ AAI systems can support industry access to research infrastructure

Creation of European ‘Superfacilities’?

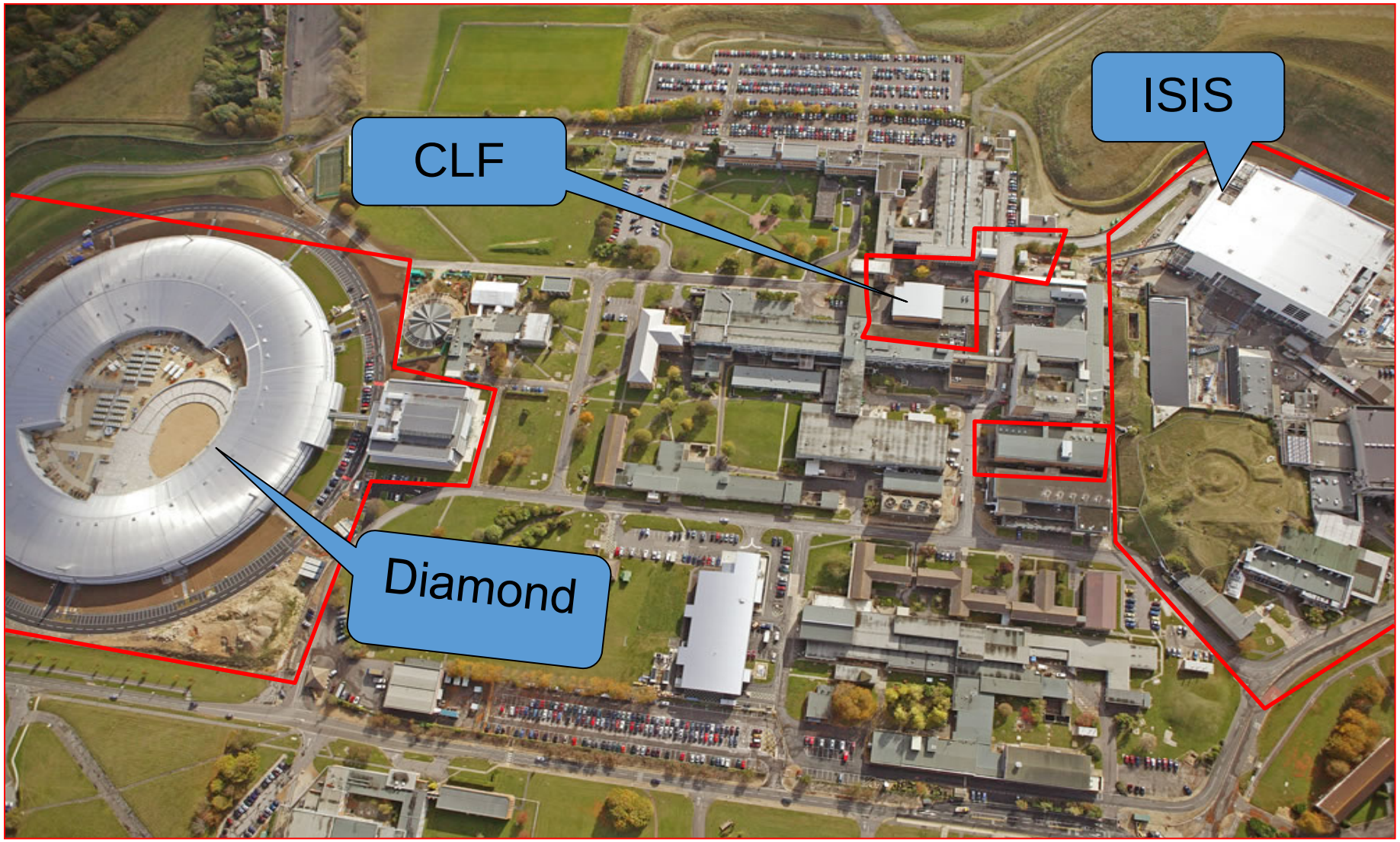
- In the US large experimental facilities are creating ‘superfacilities’ to solve advanced science questions by tightly coupling distributed resources

- 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



STFC Harwell Site Experimental Facilities in UK

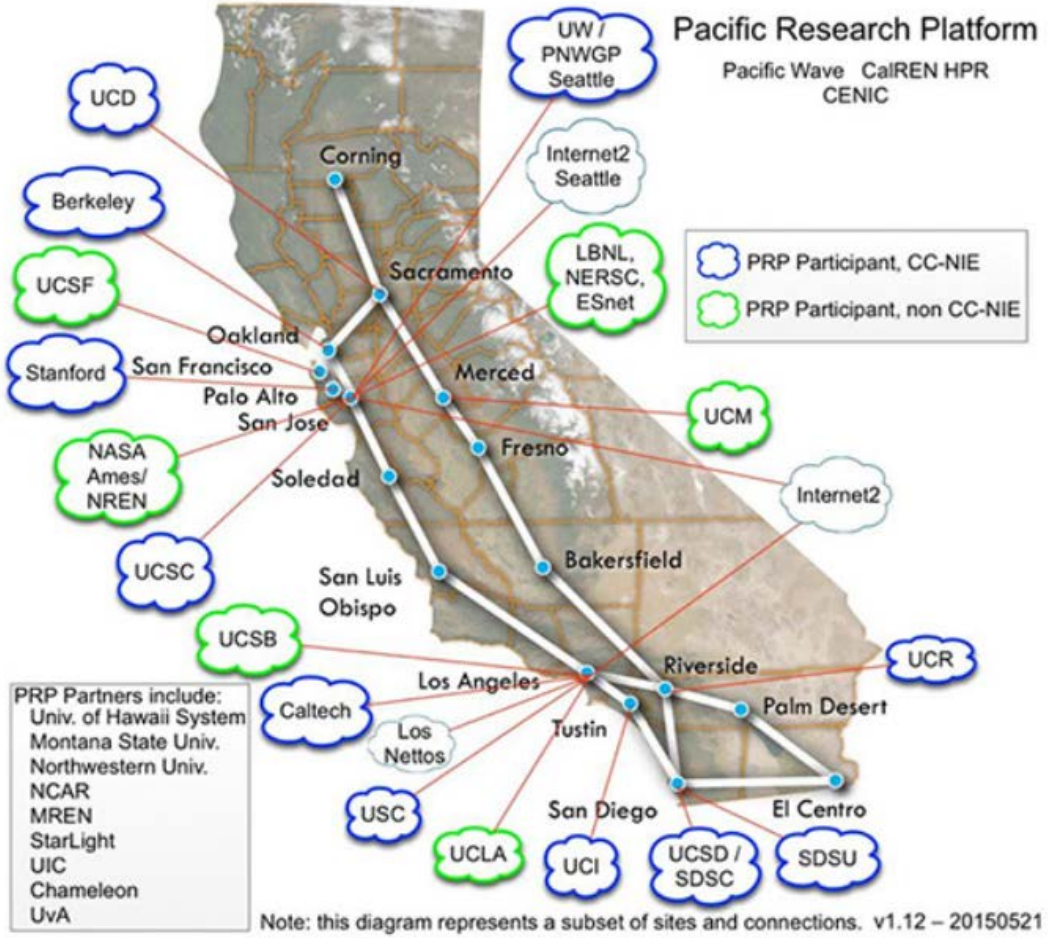


Pacific Research Platform

- 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.

Industry, Data-Scientists and e-Infrastructure

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

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

Plus Supercomputing and HPC resources

Added additional element in 2014

Training of Scientific Software Engineers and Data Scientists

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.

What is a Data Scientist?

Data Engineer

People who are expert at

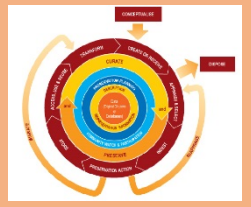
- Operating at low levels close to the data, write code that manipulates
- They may have some machine learning background.
- Large companies may have teams of them in-house or they may look to third party specialists to do the work.

Data Analyst

People who explore data through statistical and analytical methods

- They may know programming; May be an spreadsheet wizard.
- Either way, they can build models based on low-level data.
- They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.

Data Steward



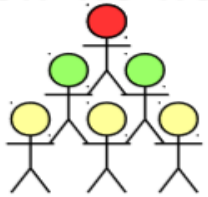
People who think to managing, curating, and preserving data.

- They are information specialists, archivists, librarians and compliance officers.
- This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

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

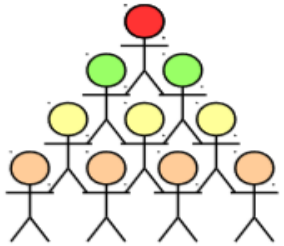
Scientist career paths?

How we worked



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

How we work



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

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

Slide thanks to Bryan Lawrence

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.’

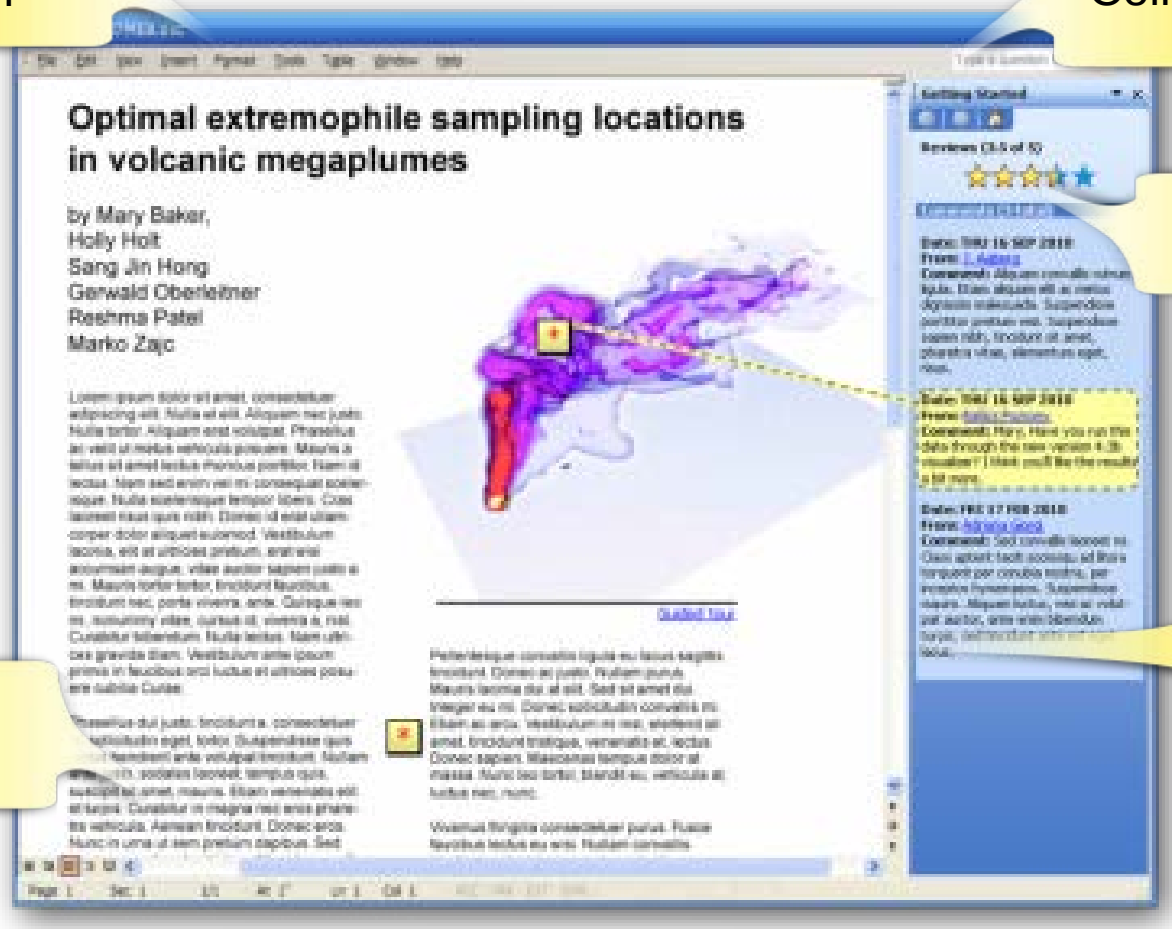
Michael Lesk on Just-in-time instead of Just-in-case?

‘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. ‘

Vision for a New Era of Research Reporting

Reproducible Research

Collaboration

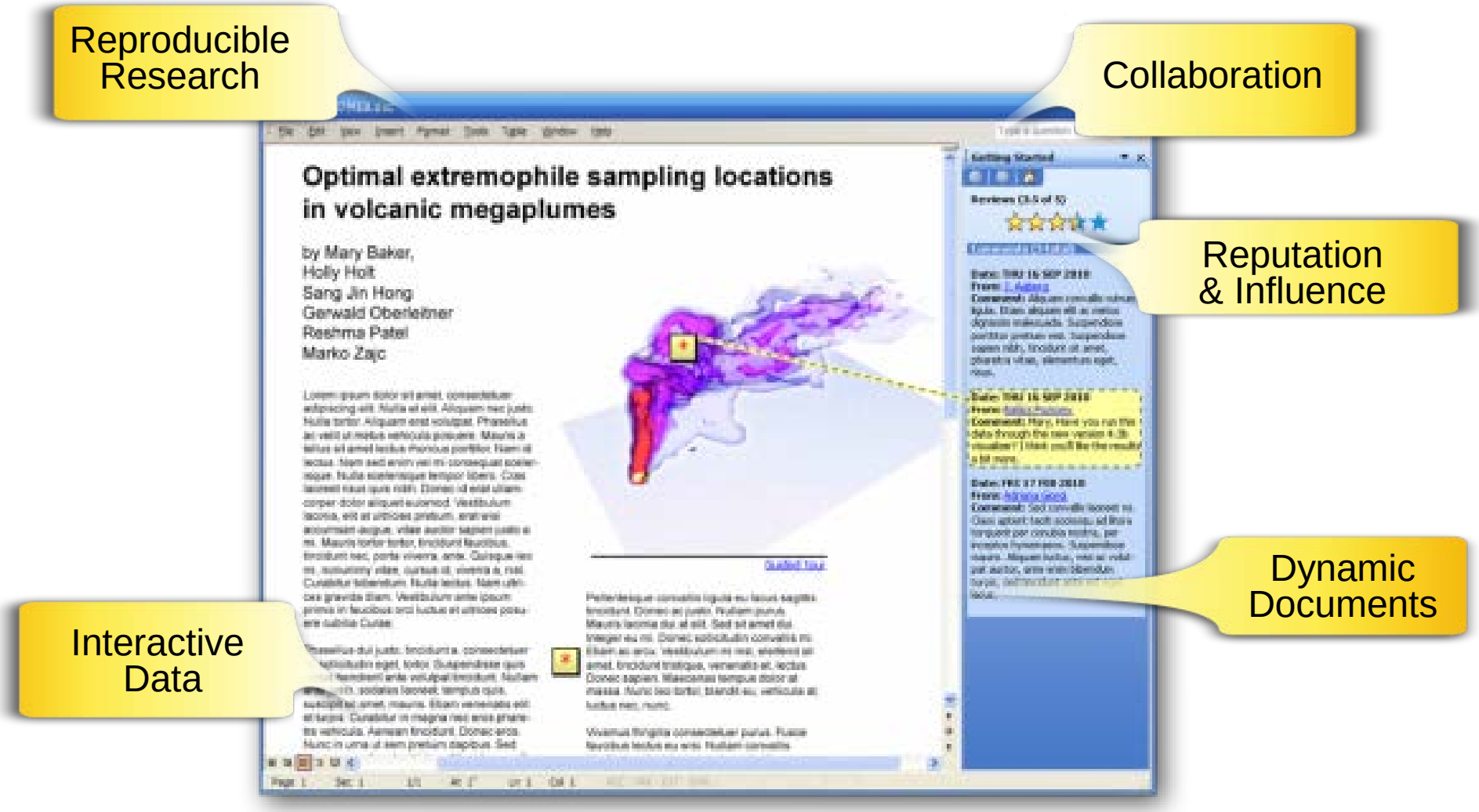


Reputation & Influence

Dynamic Documents

Interactive Data

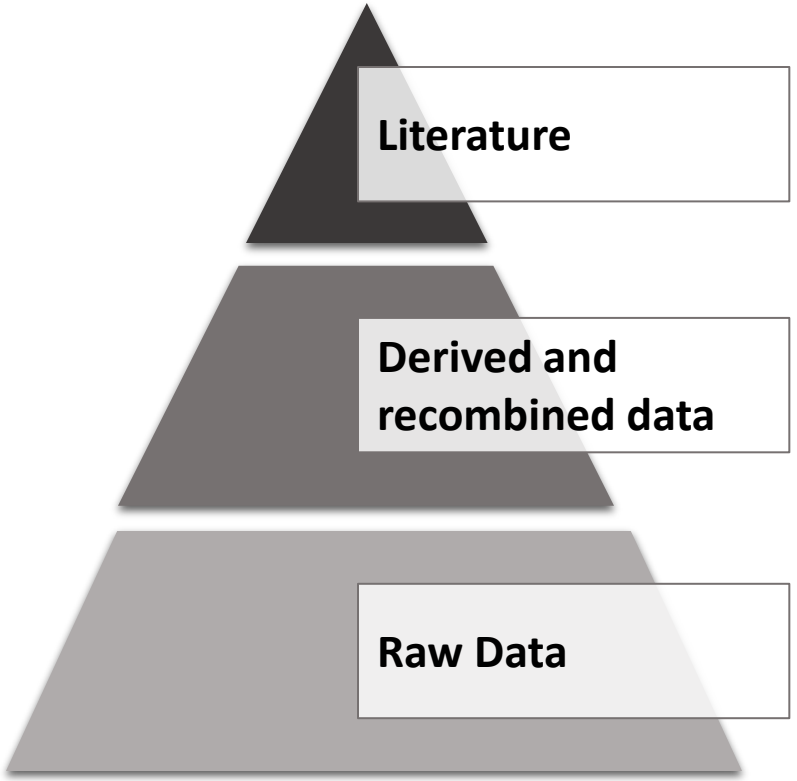
Vision for a New Era of Research Reporting



Thanks to Bill Gates SC05

Jim Gray's Vision: All Scientific Data Online

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



From Jim Gray's last talk