Data Science in the Physical Sciences

Tony Hey

Chief Data Scientist

Science and Technology Facilities Council

Harwell

bing "data scientist"

Web Images Videos Maps News Explore

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

2,160,000 RESULTS Any time

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.

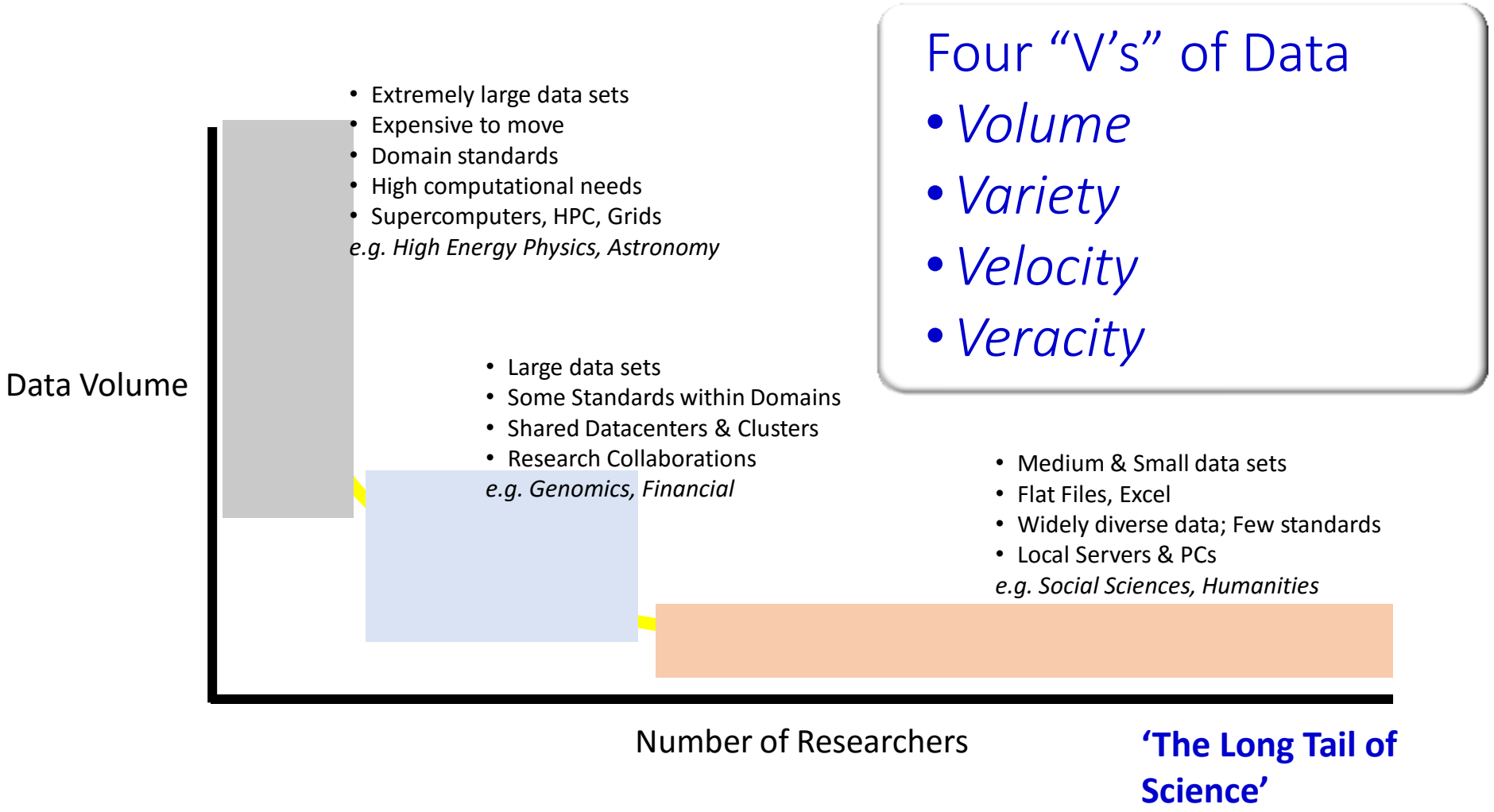
What is a Data Scientist?

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 	**People who think to managing, curating, and preserving data.** • They are information specialists, archivists, librarians and compliance officers. • This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

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

The Data Deluge – Data-Intensive Science

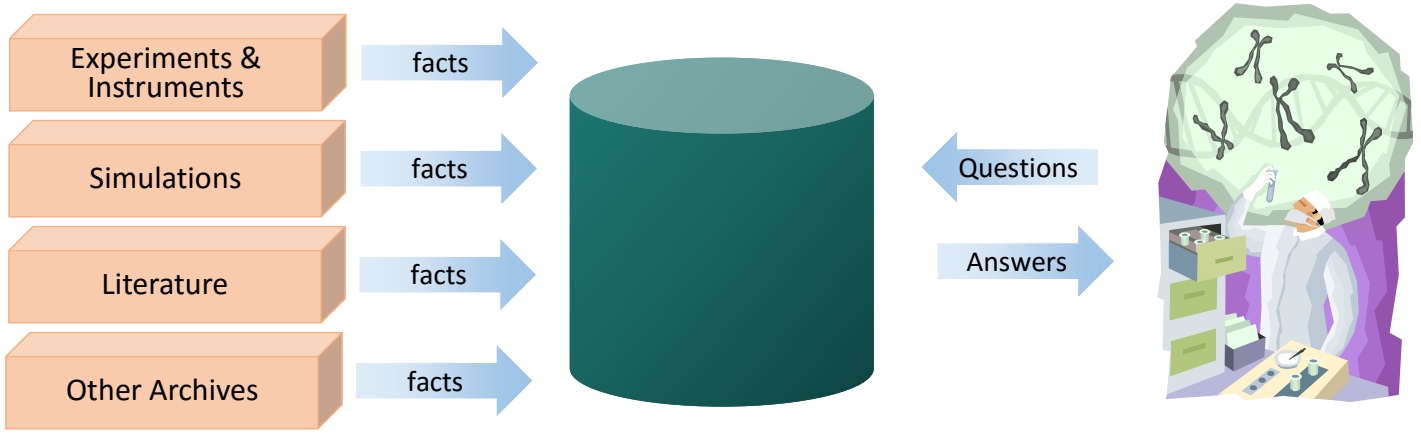
Much of Science is now Data-Intensive



Jim Gray, Turing Award Winner

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

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…

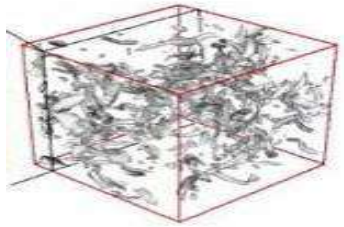
Last few decades – **Computational Science**

- 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 from Large Facilities

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 hits in 6 years
- 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

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Large Synoptic Survey Telescope

A revolutionary 3.2 gigapixel camera mounted in a massive ground-based telescope will produce unprecedented views of the cosmos, driving discoveries with the widest, densest, and most complete images of our universe ever captured.

New Visions

The Large Synoptic Survey Telescope (LSST) will peer into space as no other telescope can. This new facility will create an unparalleled wide-field astronomical survey of our universe – wider and deeper in volume than all previous telescopes combined. The combination of a 3,200 megapixel camera sensor array, a powerful supercomputer, a sophisticated data processing and distribution network, and a massive telescope stationed on a mountaintop in Chile promises to cast light on mysteries fundamental to our understanding of the universe. From the distant signatures of dark energy to the dangers of near-earth asteroids, LSST will capture it all. Three central considerations dictated the design of LSST: that it be wide, fast, and deep.

WIDE A 3.5-degree field of view will produce a complete survey of the entire night sky every three days, an impossible task for a smaller device.

FAST 15-second exposures will chart the changes in space with 800 panoramic shots a night, tracking phenomena from the redshift at the edge of the universe to fast-moving asteroids closer to home.

DEEP The sensitive camera array will ultimately image five billion galaxies, catching the subtle glow of the far-reaches of space in unprecedented detail.

Brookhaven LSST News

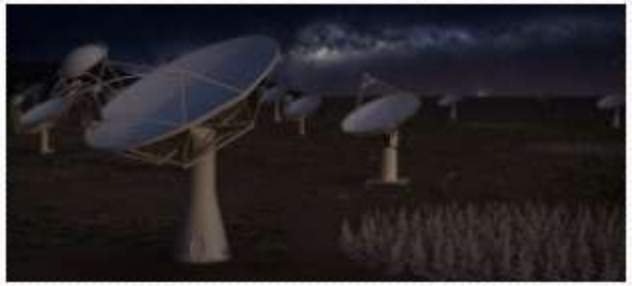
World's Most Powerful Digital Camera Sees Construction Green Light

Galaxy-Gazing Telescope Sensors Pass Important Vision Tests

503rd Brookhaven Lecture Wednesday, 3/25: Looking for Light From Dark Energy With the LSST

World's Most Powerful Camera Receives Funding Approval

SKA Project



Artist impression of the Square Kilometre Array

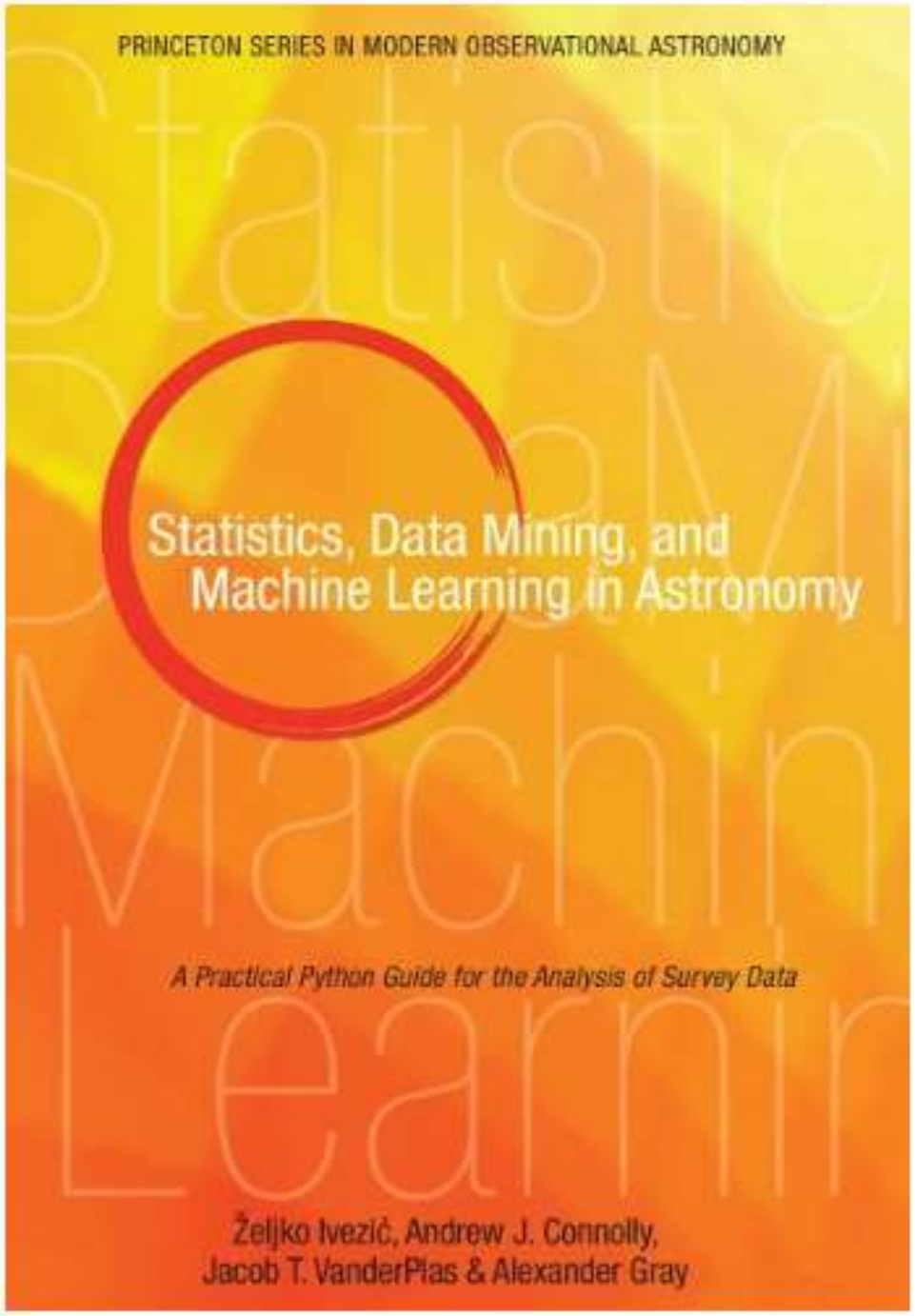
The Square Kilometre Array (SKA) project is an international effort to build the world's largest radio telescope, with eventually over a square kilometre (one million square metres) of collecting area. The scale of the SKA represents a huge leap forward in both engineering and research & development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

Latest News

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

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

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



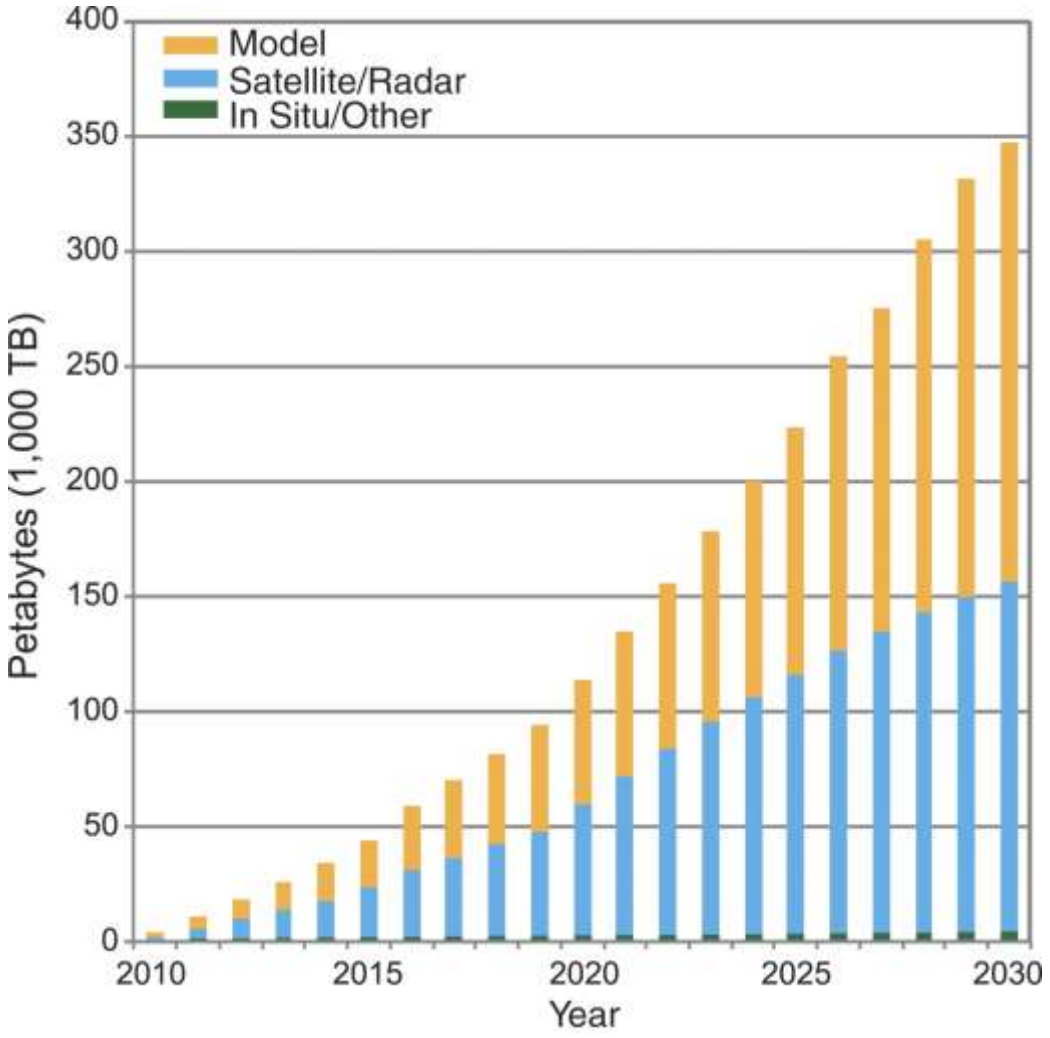
This book provides a comprehensive and accessible introduction to the cutting-edge statistical methods needed to efficiently analyze complex data sets from astronomical surveys such as the Panoramic Survey Telescope and Rapid Response System, the Dark Energy Survey, and the upcoming Large Synoptic Survey Telescope.

CEDA: Centre for Environmental Data Analysis

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

Growth of Climate Data

Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you are not a climate scientist.



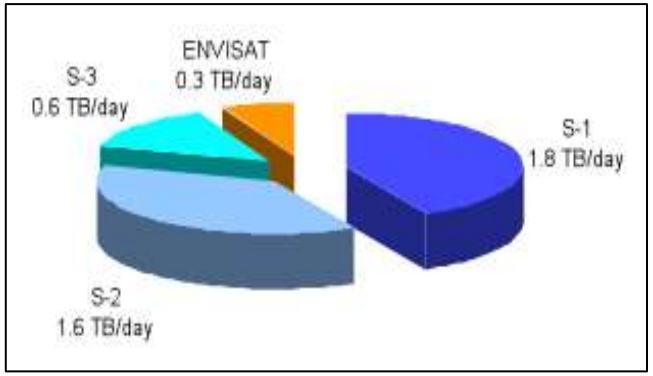
Jonathan T. Overpeck et al. Science 2011;331:700-702

Science
AAAS

Published by AAAS

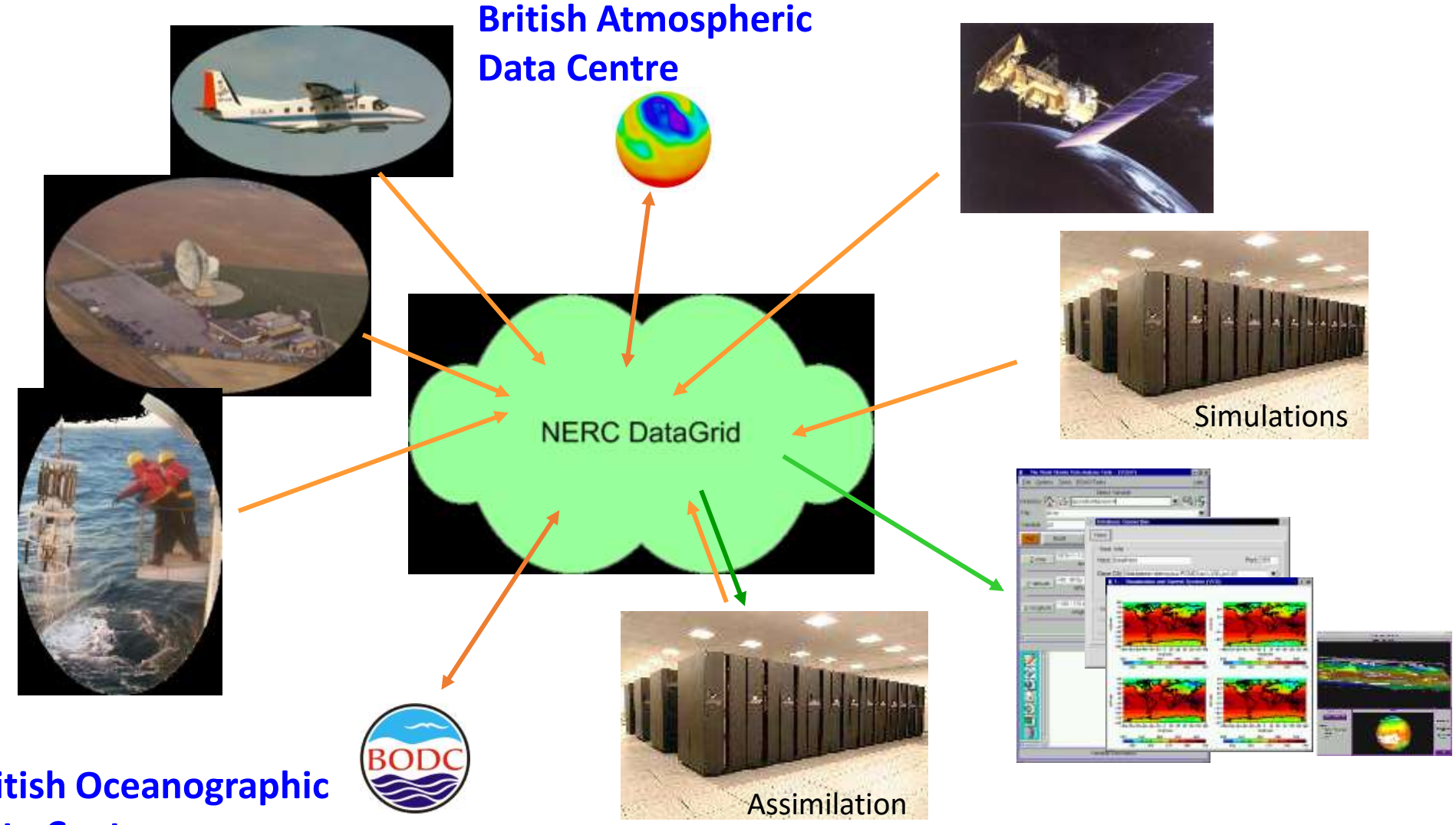
Copernicus Sentinel Satellites

Launch dates:
S1A 3rd April 2014
S2A 23rd June 2015
S3A 10th Dec 2015
S5P ~ June 2016



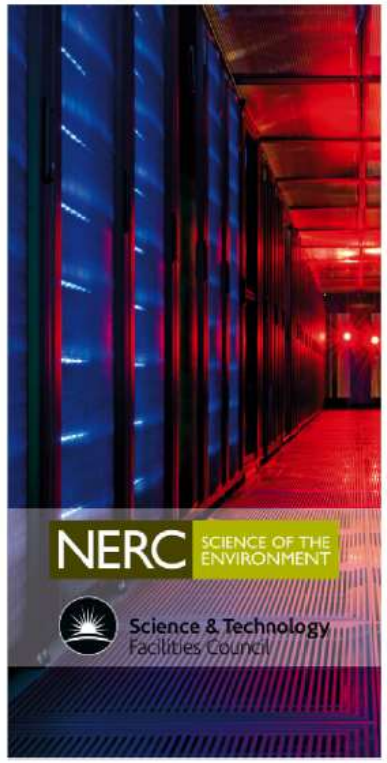
	Daily data rates	Product archive/year
Sentinel 1A, 1B	1.8 TB/day raw data	2 PB/year
Sentinel 2A	1.6 TB/day raw data	2.4 PB/year
Sentinel 3A	0.6 TB/day raw data	2 PB/year

The NERC DataGrid Project



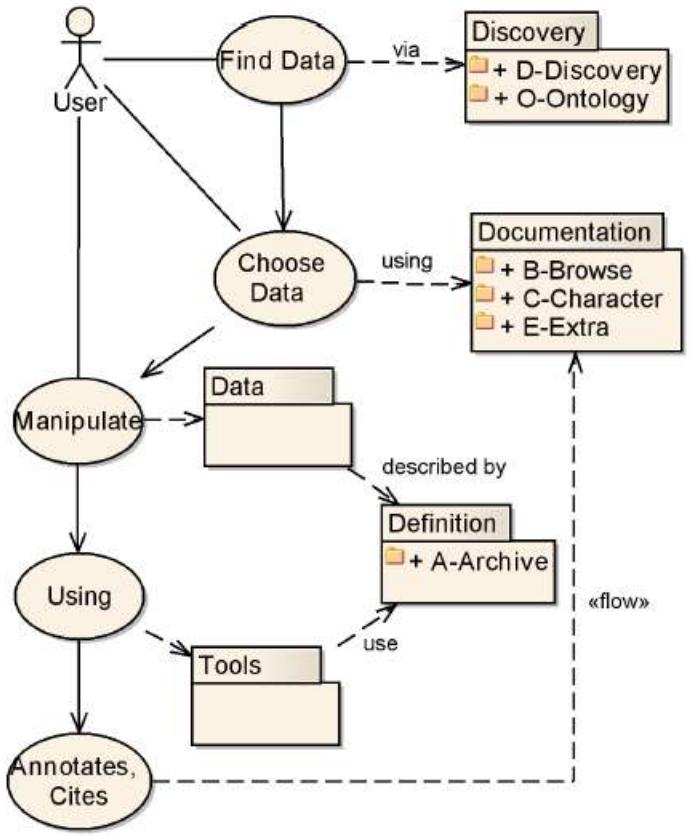
JASMIN infrastructure

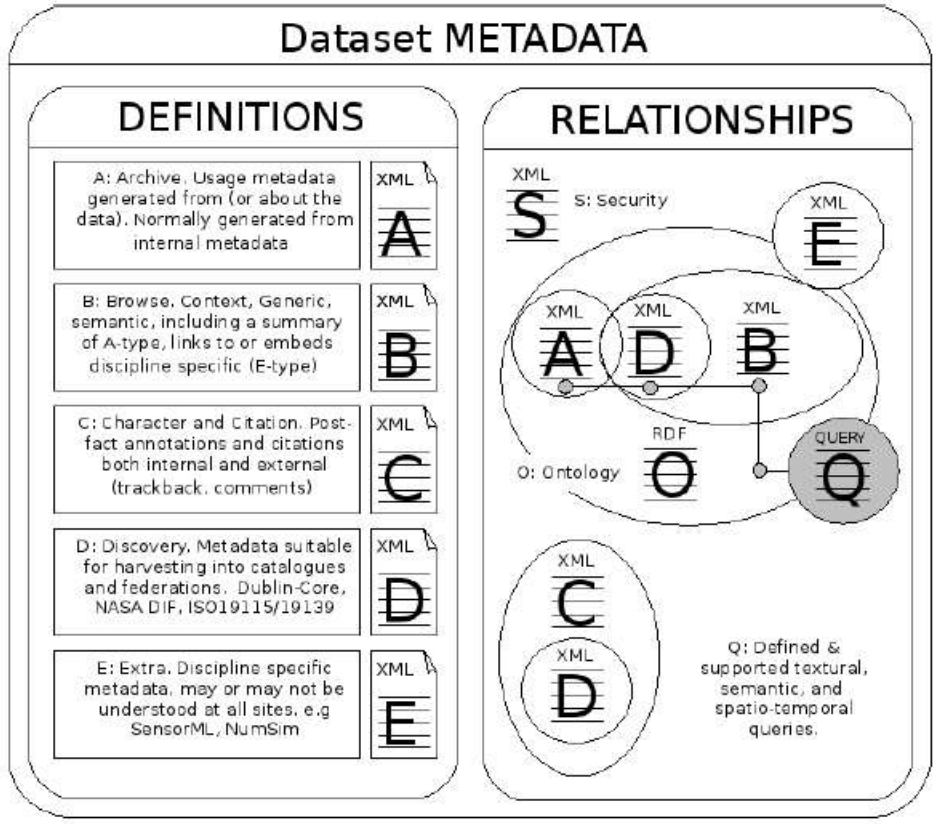
Part data store, part supercomputer, part private cloud…



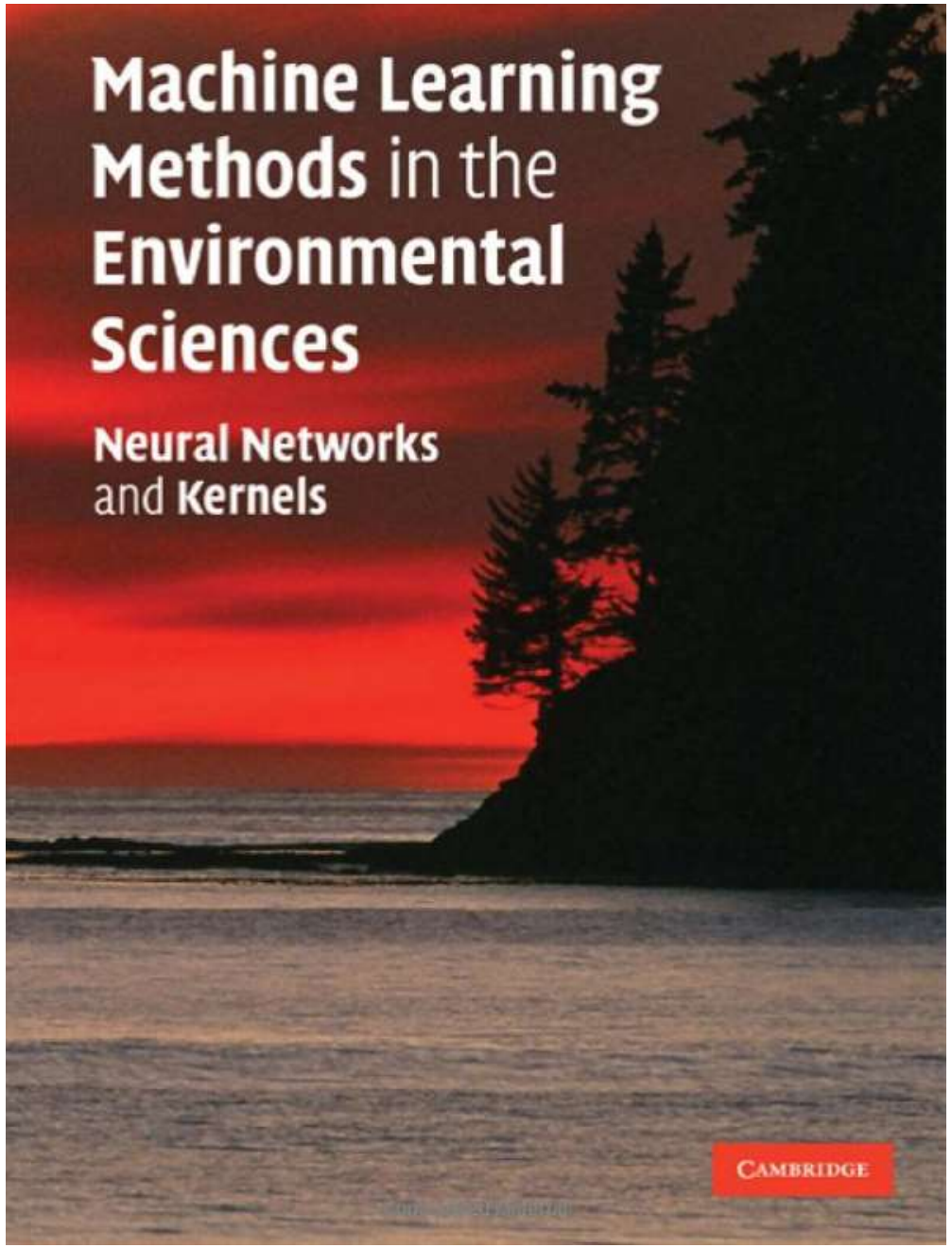
- 16 PB Fast Storage (Panasas, many Tbit/s bandwidth)
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.
- Some high memory nodes, a range, bottom heavy.

Information and Workflow





Lawrence, Lowry, Miller, Snaith and Woolf: Information in environmental data grids, Phil. Trans. R. Soc. A (2009) doi:10.1098/rsta.2008.0237



Due to their powerful nonlinear modeling capability, machine learning methods today are used in satellite data processing, general circulation models(GCM), weather and climate prediction, air quality forecasting, analysis and modeling of environmental data, oceanographic and hydrological forecasting, ecological modeling, and monitoring of snow, ice and forests.

GridPP – UK Computing for Particle Physics

GridPP prototype Grid

> 1,000 CPUs

- 500 CPUs at the Tier-1 at RAL
- > 500 CPUs at 11 sites across UK organised in 4 Regional Tier-2s

> 500 TB of storage

> 800 simultaneous jobs

Integrated with international LHC Computing Grid (LCG)

> 5,000 CPUs

> 4,000 TB of storage

> 70 sites around the world

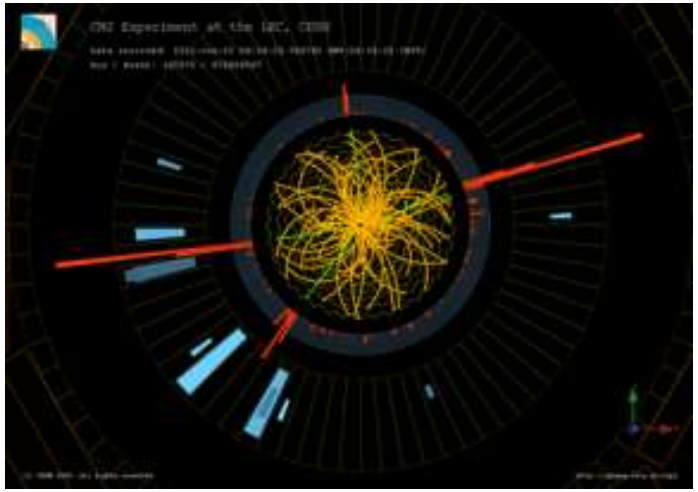
> 4,000 simultaneous jobs

- monitored via Grid Operations Centre (RAL)

From UK AHM 2004 Presentation

The GRIDPP Tier-1 for the LHC

- High capacity service for hunting Higgs Particle
- Remote job submission (Grid)
- Low cost commodity hardware
- Open source middleware
- Delivered 99.2% availability in 2013 (70 hours lost all causes)
- Moves 50 petabytes/year globally (wide area transfers)
- Moves 300 petabytes/year internally (data processing)
- Run 2 now underway(data rates and volumes double)



- 13.6 PB disk
- 16 PB tape
- 10,000 CPU cores
- 2000 servers
- 40Gb network
- 10Gb/s direct optical link to CERN

Completed • $13,000 • 1,785 teams

Higgs Boson Machine Learning Challenge

Mon 12 May 2014 – Mon 15 Sep 2014 (16 months ago)

Dashboard

Home
Data
Make a submission

Information
Description
Evaluation
Rules
Prizes
About the Sponsors
Timeline
Winners

Forum

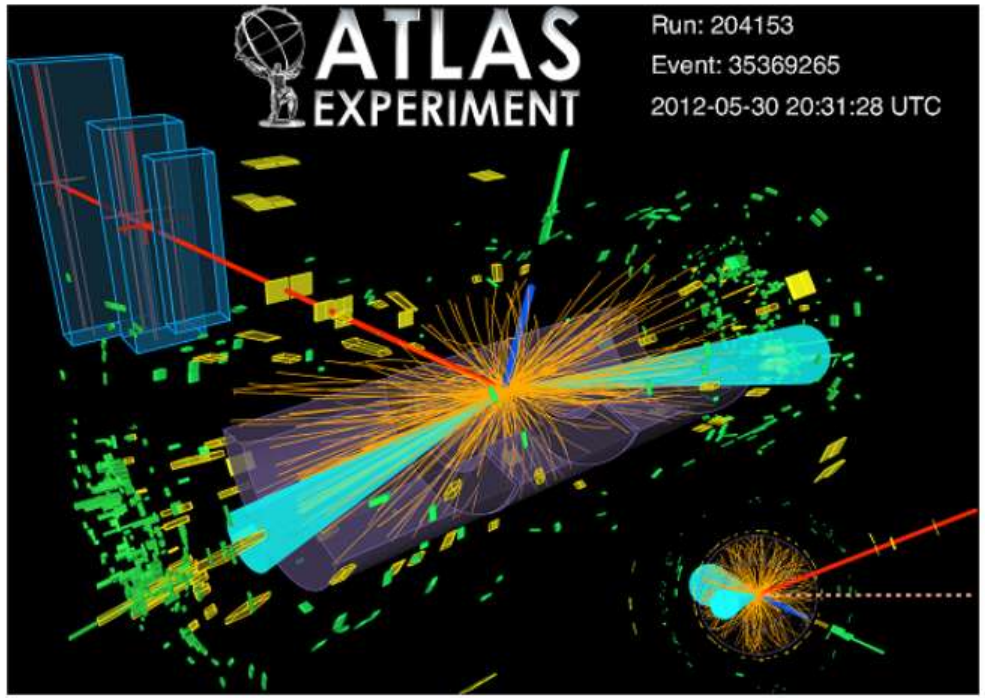
Leaderboard
Public
Private

Leaderboard

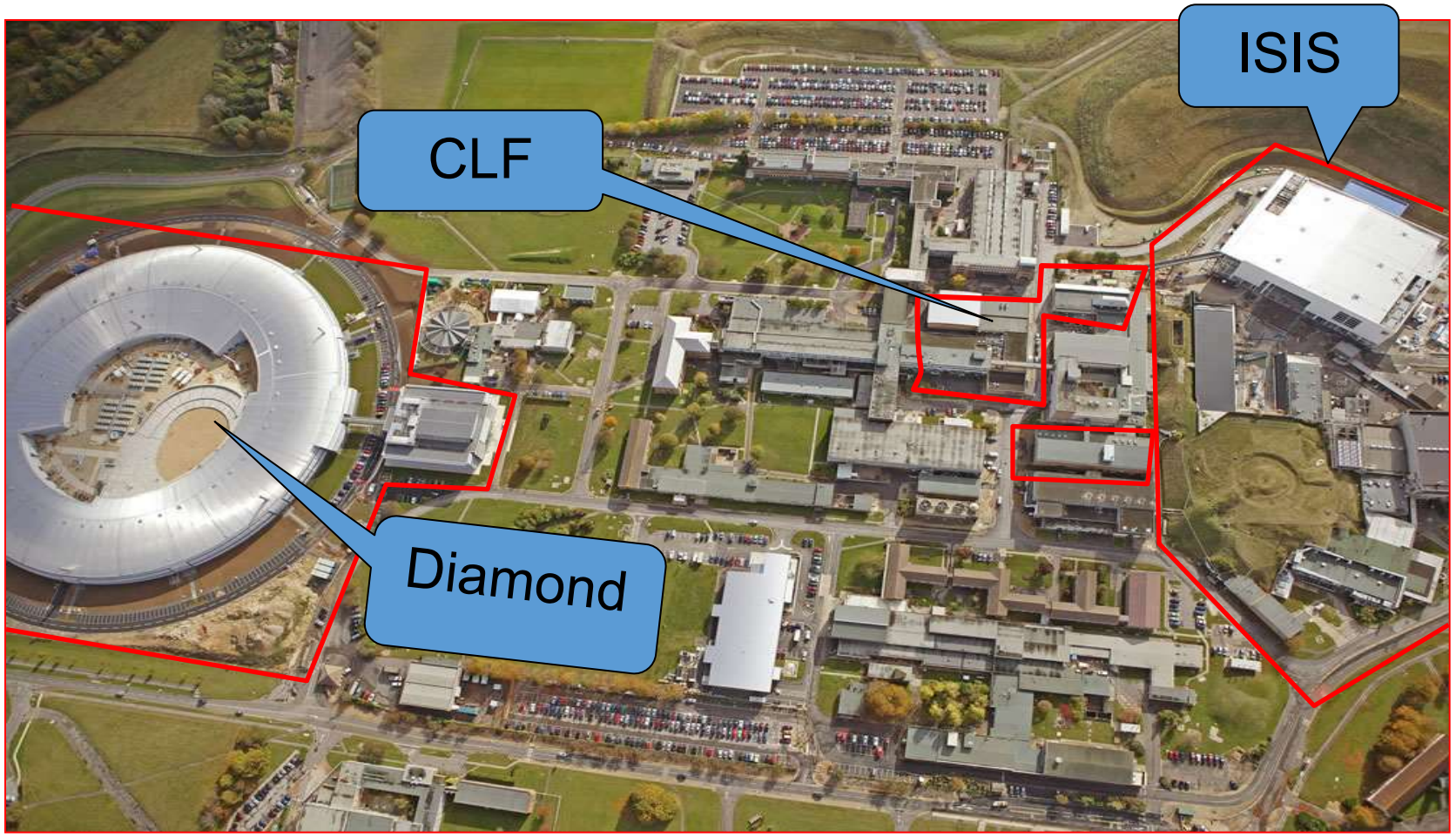
1. Gábor Melis
2. Tim Salimans
3. nhlx5haze
4. ChoKo Team
5. cheng chen
6. quantify
7. Stanislav Semenov & Co (HSE Yandex)

Competition Details » Get the Data » Make a submission

Use the ATLAS experiment to identify the Higgs boson



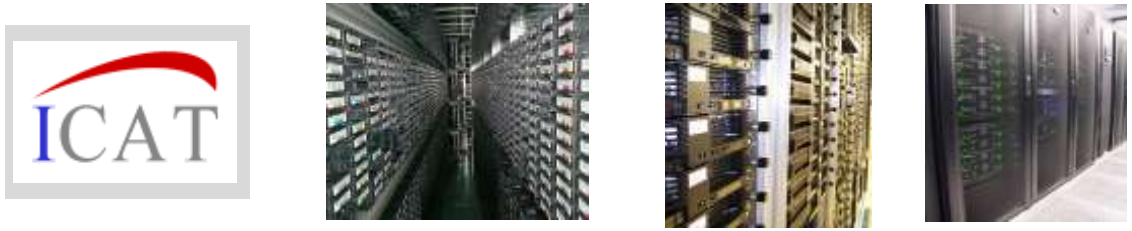
Harwell Site Experimental Facilities



Applications of Synchrotron Radiation
Rolls-Royce Trent 800
Normal Conformer
Rogue Conformer
(speculative)

In- and Post-experimental support

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



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

Infrastructure + Software + Expertise

ISIS:IMAT

DLS:I12/I13

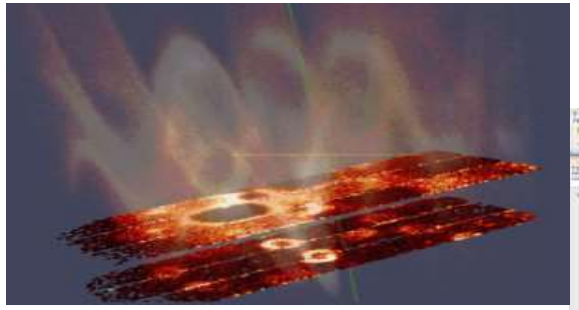
- **Tomography**: Dealing with high data volumes – 200Gb/scan, ~5 TB/day (one experiment at DLS)
- **MX**: high data volumes, smaller files, but a lot more experiments
- Hard to move the data – needs to be handled at the facility?

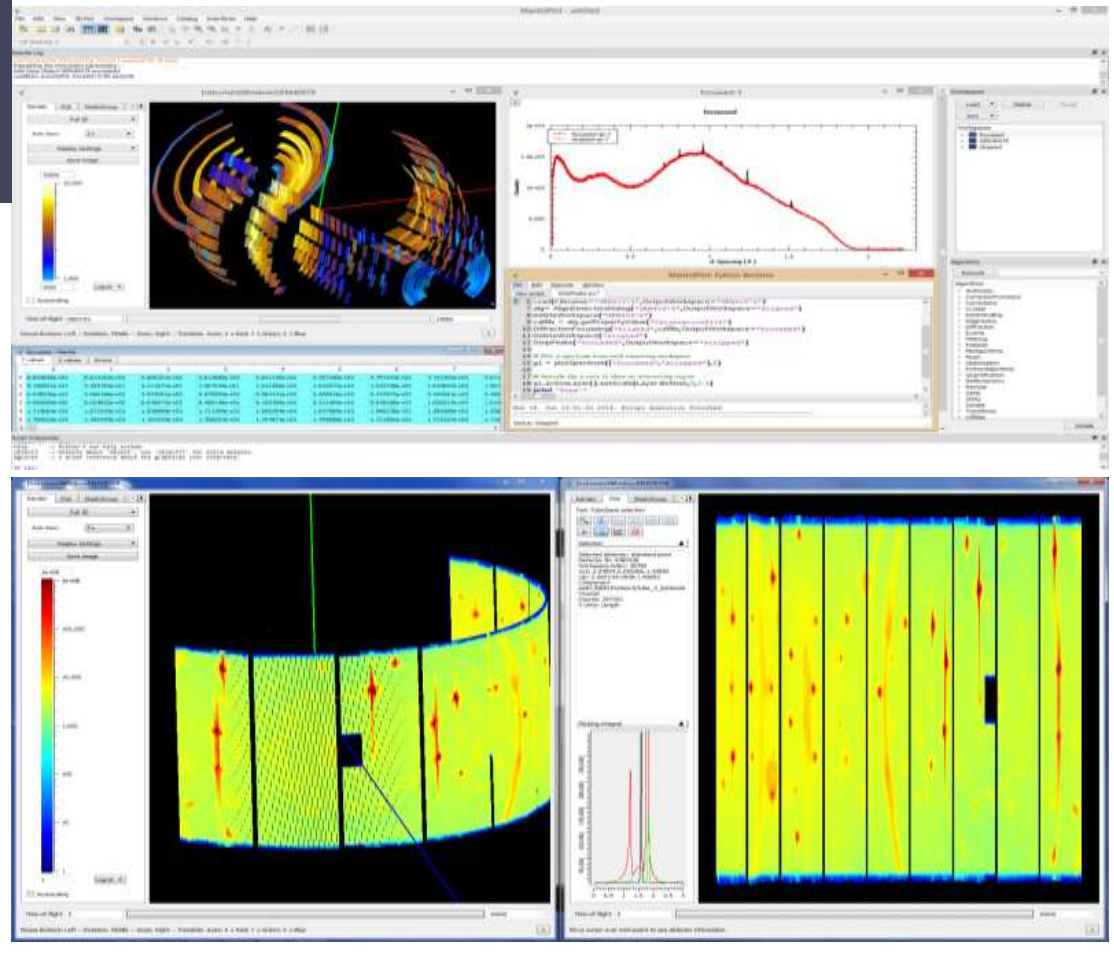
Erica Yang, Sri Nagella

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

ISIS Data Reduction Stage

The Mantid Project





- Aims
  - Data reduction/analysis framework for neutron scattering and muon experiments
  - Without restricting the needs of the instrument scientists
    - Cross-platform
    - Easily extensible
    - Freely redistributable
    - Open source
- Adoption
  - 3 Partner Facilities
  - 9 Contributing Organisation
  - In use on 52 instruments worldwide

Science & Technology Facilities Council
ISIS

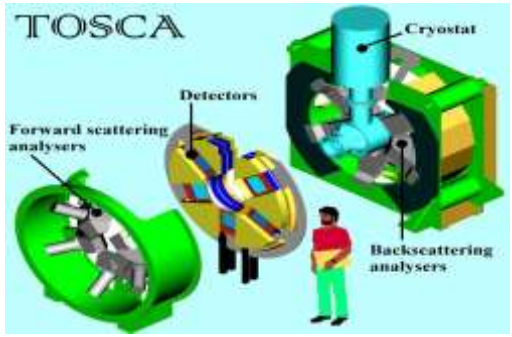
OAK RIDGE National Laboratory
HIGH FLUX ISOTOPE REACTOR
SPALLATION NEUTRON SOURCE

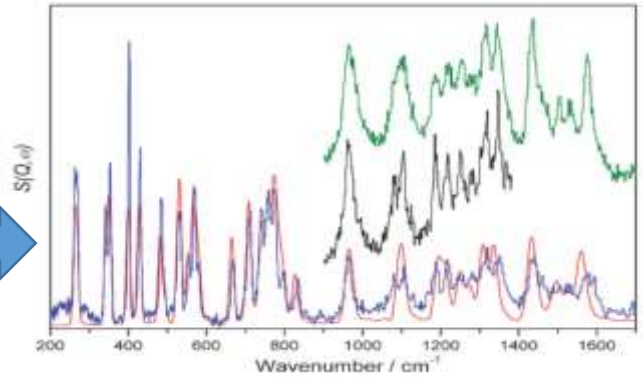
ess EUROPEAN SPALLATION SOURCE

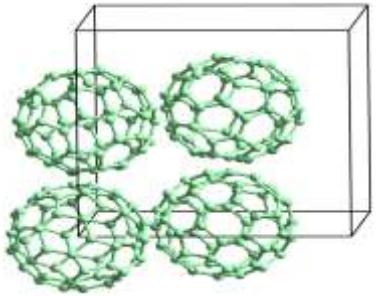
Tessella
Technology & Consulting

Ab initio modelling of ISIS data

- **Inelastic Neutron Spectroscopy (INS) with TOSCA instrument on ISIS**
- **CASTEP density-functional theory code running on SCARF HPC cluster used to model spectra**
- **Results of simulation used to interpret peaks in measured INS spectrum.**
- **Can only understand experiment with aid of simulation.**







CASTEP

SCARF

Science & Technology Facilities Council
ISIS

Revealing the value of open data

Hartree Centre
Science & Technology Facilities Council

The big data analytics capabilities of the Hartree Centre and its partner IBM are enabling Democrata to help construction firms predict the presence of archaeological remains.

Work with us

We collaborate with industrial

Open Science and the US OSTP Memo

US White House Memorandum

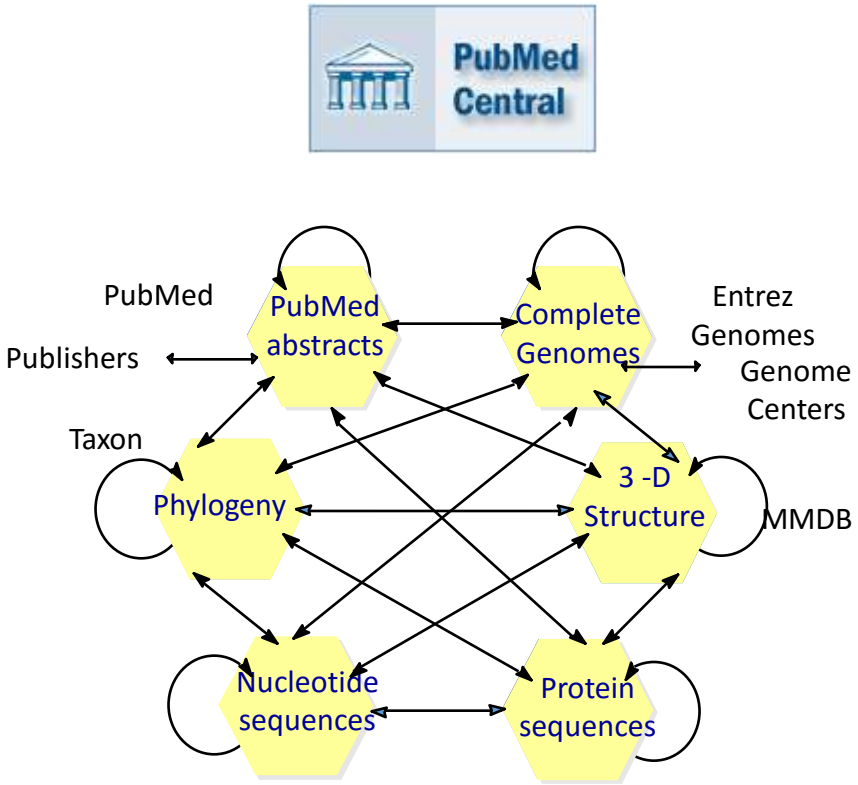
- Directive requiring the major Federal Funding agencies *"to develop a plan to support increased public access to the results of research funded by the Federal Government."*

- 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

The US National Library of Medicine

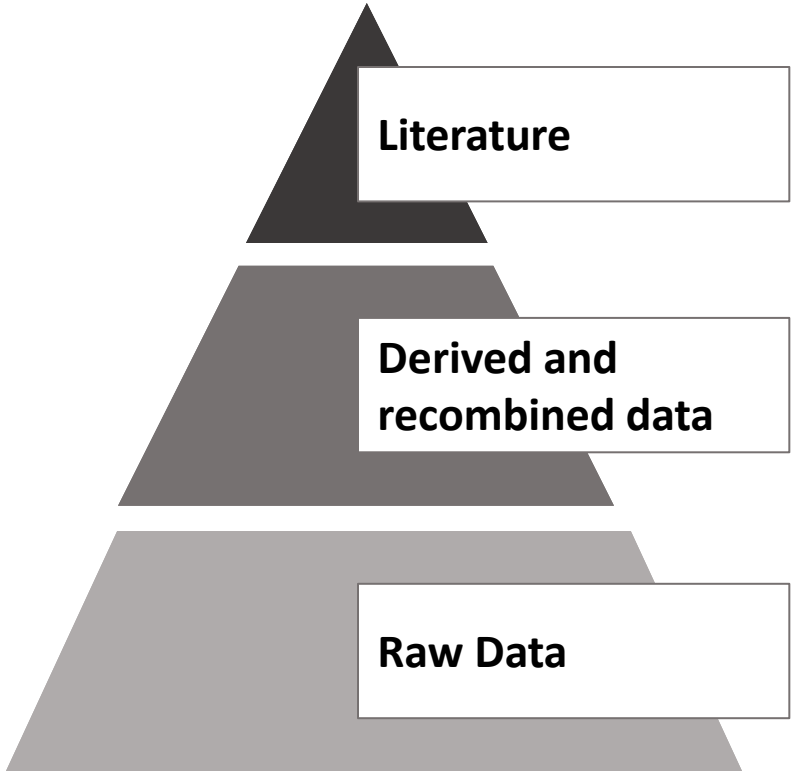
- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- Policy requires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



Entrez cross-database search

Jim Gray's Vision: All Scientific Data Online

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



(*From Jim Gray's last talk)*

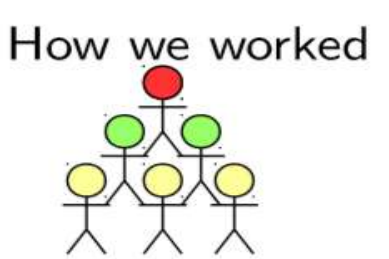
Data Science in the Future?

Digital Curation Centre

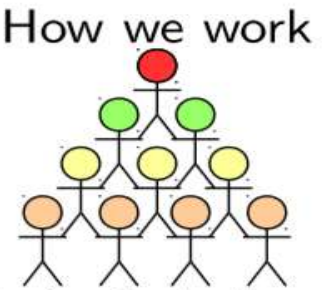
Established in 2004 with mission to work with librarians, scientists and computer scientists to examine all aspects of research data curation and preservation

- Actions needed to maintain and utilise digital data and research results over entire life-cycle
  - For current and future generations of users
- Digital Preservation
  - Long-run technological/legal accessibility and usability
- Data curation in science
  - Maintenance of body of trusted data to represent current state of knowledge in area of research
- Research in tools and technologies
  - Integration, annotation, provenance, metadata, security…..

How do we work?



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



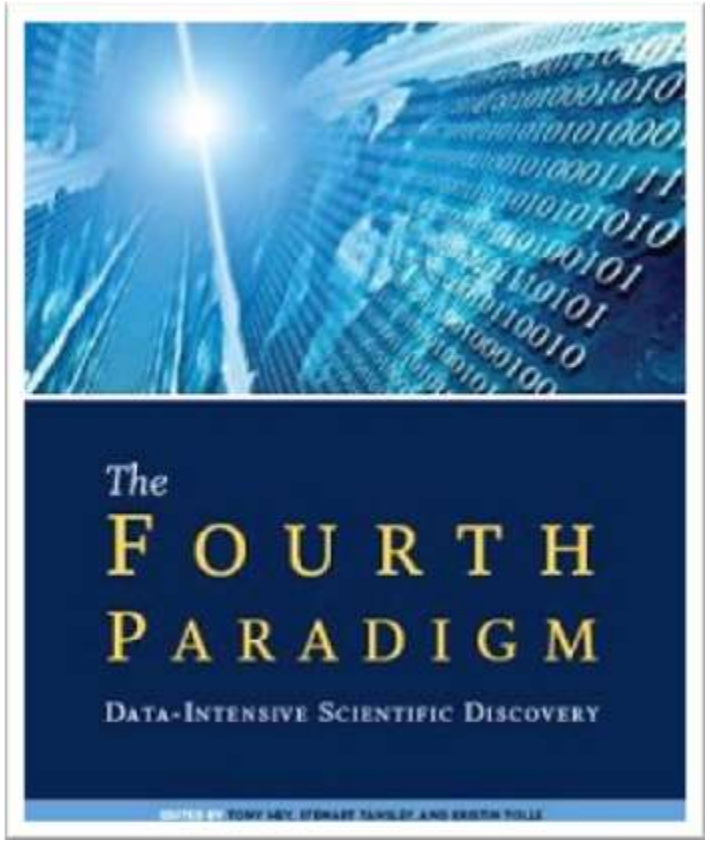
PI stands on the shoulders of her postdocs, students, software engineers and data scientists. (Are the giants down with the turtles?).

- 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

- ❑ Big Data
- ❑ Machine Learning
- ❑ Visualization
- ❑ Cloud Services

Data-Intensive Scientific Discovery



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