Science & Technology Facilities Council
Rutherford Appleton Laboratory

Data repository management in the environmental sciences in the UK

Sarah Callaghan
sarah.callaghan@stfc.ac.uk
@sorcha_ni

Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop

Boulder, CO, USA, 7-9 August 2018

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

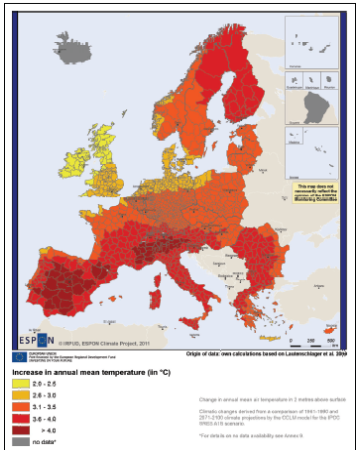
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

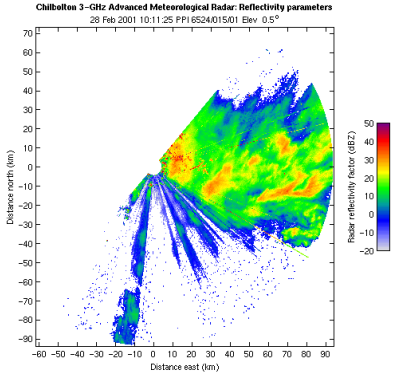
NERC Environmental Data Centres

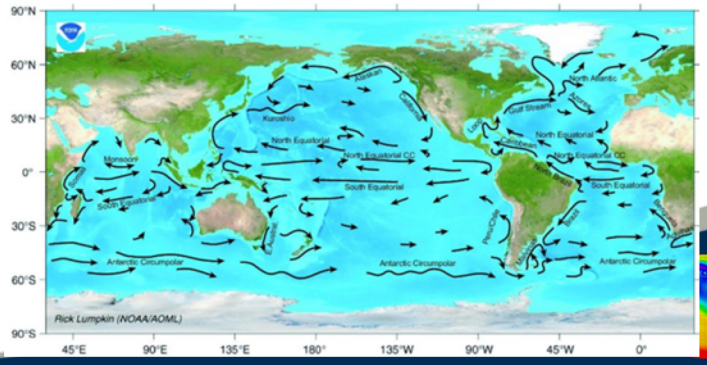
- National Geoscience Data Centre (NGDC)
  - Geoscience/ Sub-surface
- **Centre for Environmental Data Analysis (CEDA)**
  - **Atmosphere/ Solar System/ Earth Observation**
- Environmental Information Data Centre (EIDC)
  - Terrestrial / Freshwater
- Polar Data Centre (PDC)
  - Polar - regional
- British Oceanographic Data Centre (BODC)
  - Oceanographic / Marine

Data heterogeneity

1. Time series, some still being updated e.g. meteorological measurements
2. Large 4D synthesised datasets, e.g. Climate, Oceanographic, Hydrological and Numerical Weather Prediction model data generated on a supercomputer
3. 2D scans e.g. satellite data, weather radar data
4. 2D snapshots, e.g. cloud camera
5. Traces through a changing medium, e.g. radiosonde launches, aircraft flights, ocean salinity and temperature
6. Datasets consisting of data from multiple instruments as part of the same measurement campaign
7. Physical samples, e.g. fossils, ice cores







Our Mission

CEDA exists to enable the efficient delivery of environmental science programmes through effective data and information services. This is achieved by:

- operation of efficient data curation services
- facilitating environmental science by running data and information services
- development of new data service technologies
- close contact with the research community
- the development of a national capability in data management expertise
- contributing to and learning from the international community

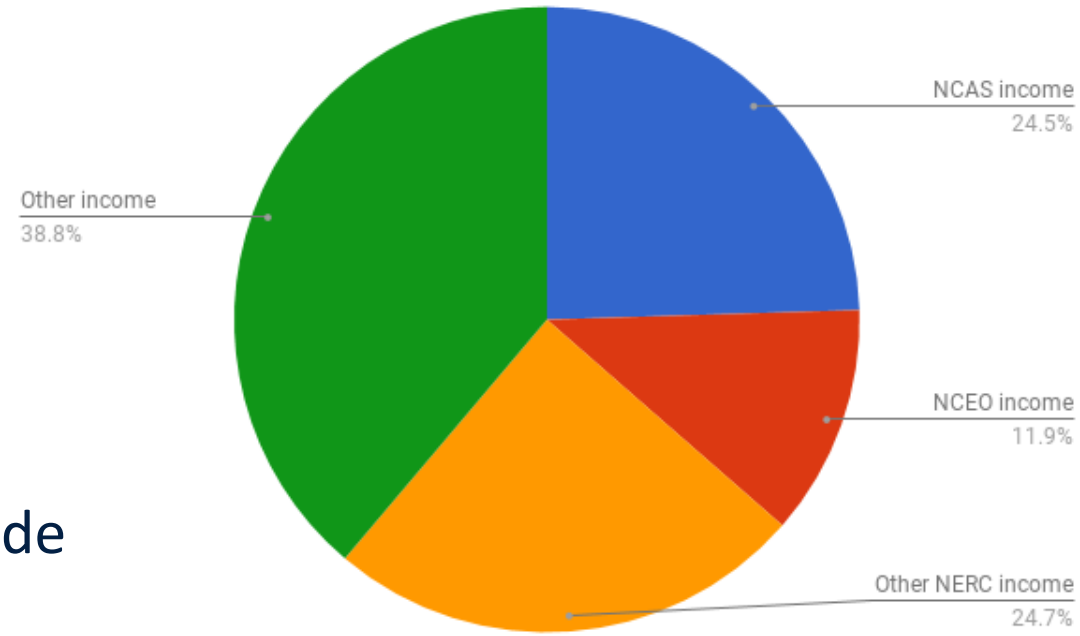
Team of ~29; mixture of Data Scientists and Software Engineers

Diverse expertise in:

- Earth observation
- Climate modelling
- Aircraft measurements
- Data standards
- … and much more!

What we do

- CEDA Data Centres
  - NCAS
  - NCEO
  - IPCC-DDC
  - UKSSDC
- JASMIN “super data cluster”
- Projects: funders include EC, ESA, UKSA, Defra, BEIS, Met Office



Tools and Services

CEDA

flight-finder.ceda.ac.uk

Apps | Bookmarks | BADC – Trac | Open Access Button | Save to Mendeley | Google Mail | EasyClock | STFC Outlook Web | Sorcha ní gCeallagh | BBC NEWS | News | Other bookmarks

Centre for Environmental Data Analysis

CEDA Archive

Search Catalogue | Get Data | Help | Tools | Deposit | News | scallagh

Dataset

GBS 20.7GHz slant path radio propagation measurements, Sparsholt site

Update Frequency:	Not Planned
Status:	Completed
Online Status:	ONLINE
Publication State:	Citable
Publication Date:	2003-09-20
DOI Publication Date:	2011-04-01
Download Stats:	last 12 months

Access Granted | Download

Abstract

The GBS (Global Broadcast Service) dataset is a series of radio attenuation measurements made at three sites in the UK: Chilbolton and Sparsholt, both in southern UK, and Dundee in Scotland. The aim of the experiment was to make long term measurements of the signal strength received from a 20.7GHz beacon on the US Department of Defense satellite UFO-9 at multiple sites, in order to determine whether the use of site diversity as a fade mitigation technique would be effective. The dataset spans a period of 3 years, from August 2003 to August 2006 with signal attenuation sampled once per second.

This dataset is cited in:
S. A. Callaghan, J. Waight, J.L.Agnew, C. J. Walden, C.L.Wrench , S. Ventouras "The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK", Geoscience Data Journal, 17 March 2013, DOI: 10.1002/gdj3.2

Research; Callaghan, S.A.; Waight, J.; Walden, C.J.; Agnew, J.; Ventouras, S. (2011): GBS 20.7GHz slant path radio propagation measurements, Sparsholt site. NCAS British Atmospheric Data Centre, *01 April 2011*. doi:10.5285/E8F43A51-0198-4323-A926-FE69225D57DD. http://dx.doi.org/10.5285/E8F43A51-0198-4323-A926-FE69225D57DD

Abbreviation: Not defined

Coverage

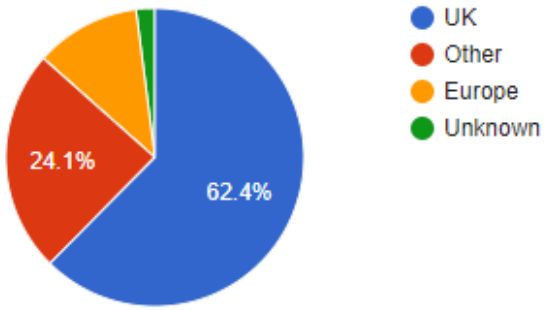
Temporal Range

Start time:	2003-10-08T00:00:00
End time:	2005-03-31T00:00:00

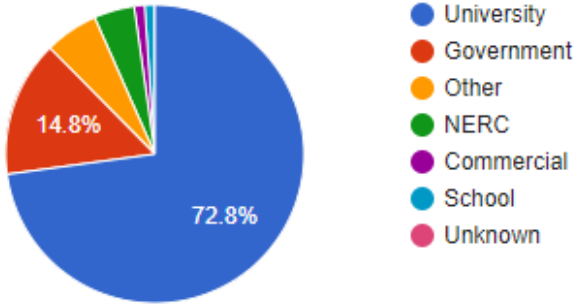
Geographic Extent

CEDA Data and Users

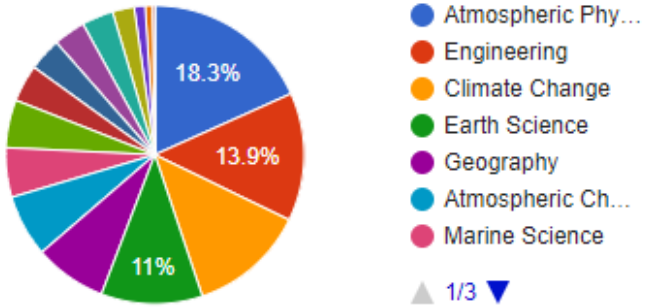
Number of users by Area

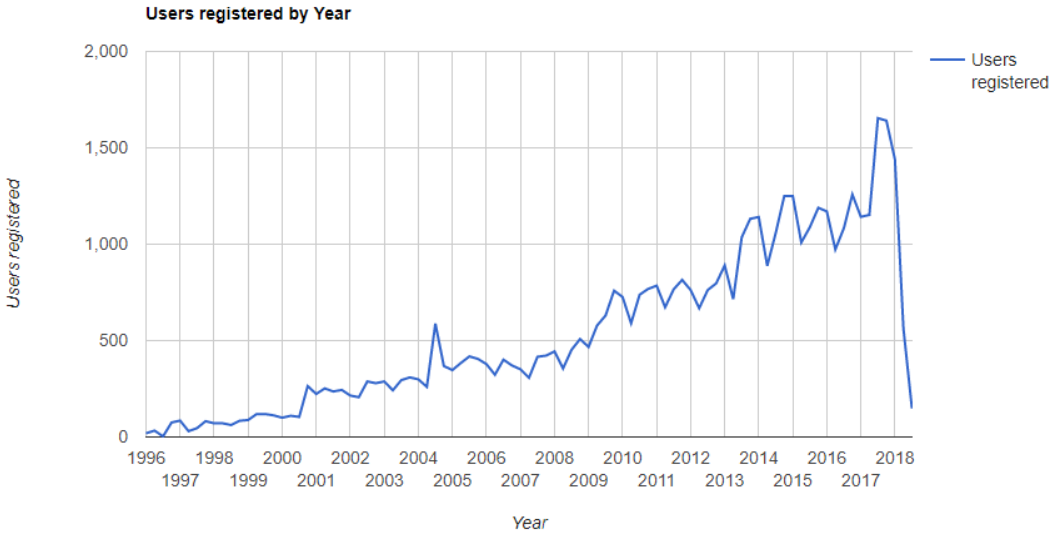


Number of users by Institute type



Number of users by Field



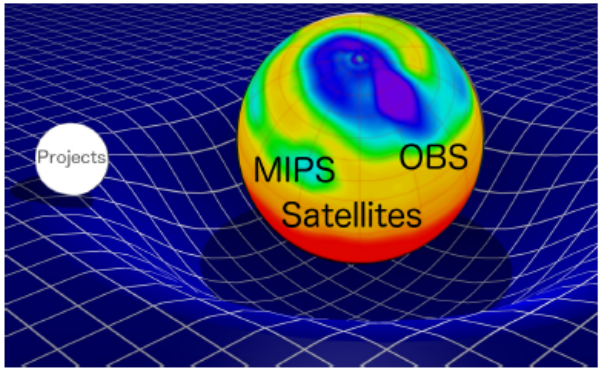


Data Type	Data Volume (Petabytes)
Earth Observation	4
Atmospheric Science	2
Total	6 PB

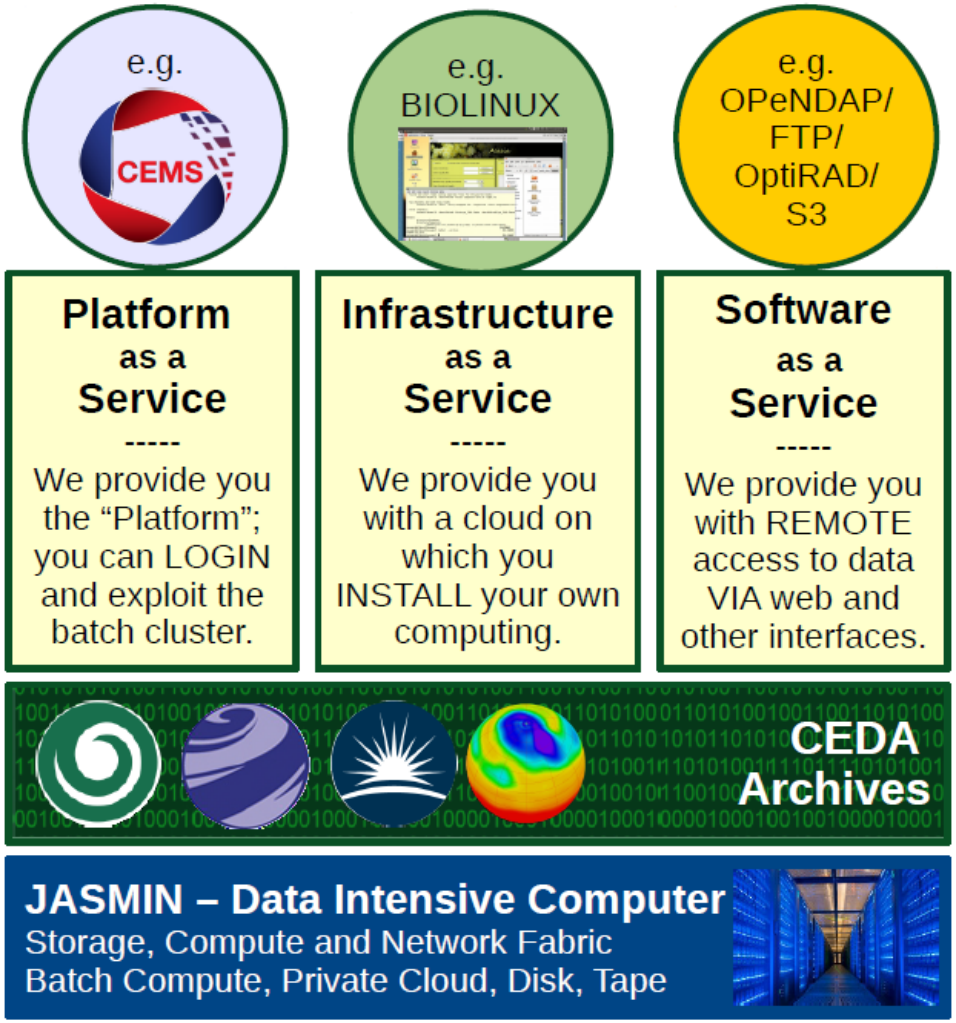
- ~ 550 datasets
- ~ 150 million files
- > 44,000 registered users

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

JASMIN – The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.



Logical View

CEDA Data Centres

IPCC Data Distribution Centre

ESGF

Other …

NERC Managed Analysis Compute

NERC Cloud Analysis Compute

Other compute …

CEDA Archive Services

Data Centres, Curation, DB systems

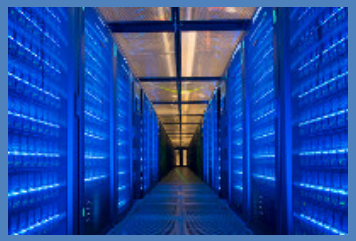
User management, External Helpdesk

Analysis Environment

Compute:

Science VMs + User Management, PaaS and IaaS Cloud,
Group Workspaces: Fast Disk & Elastic Tape

External Helpdesk

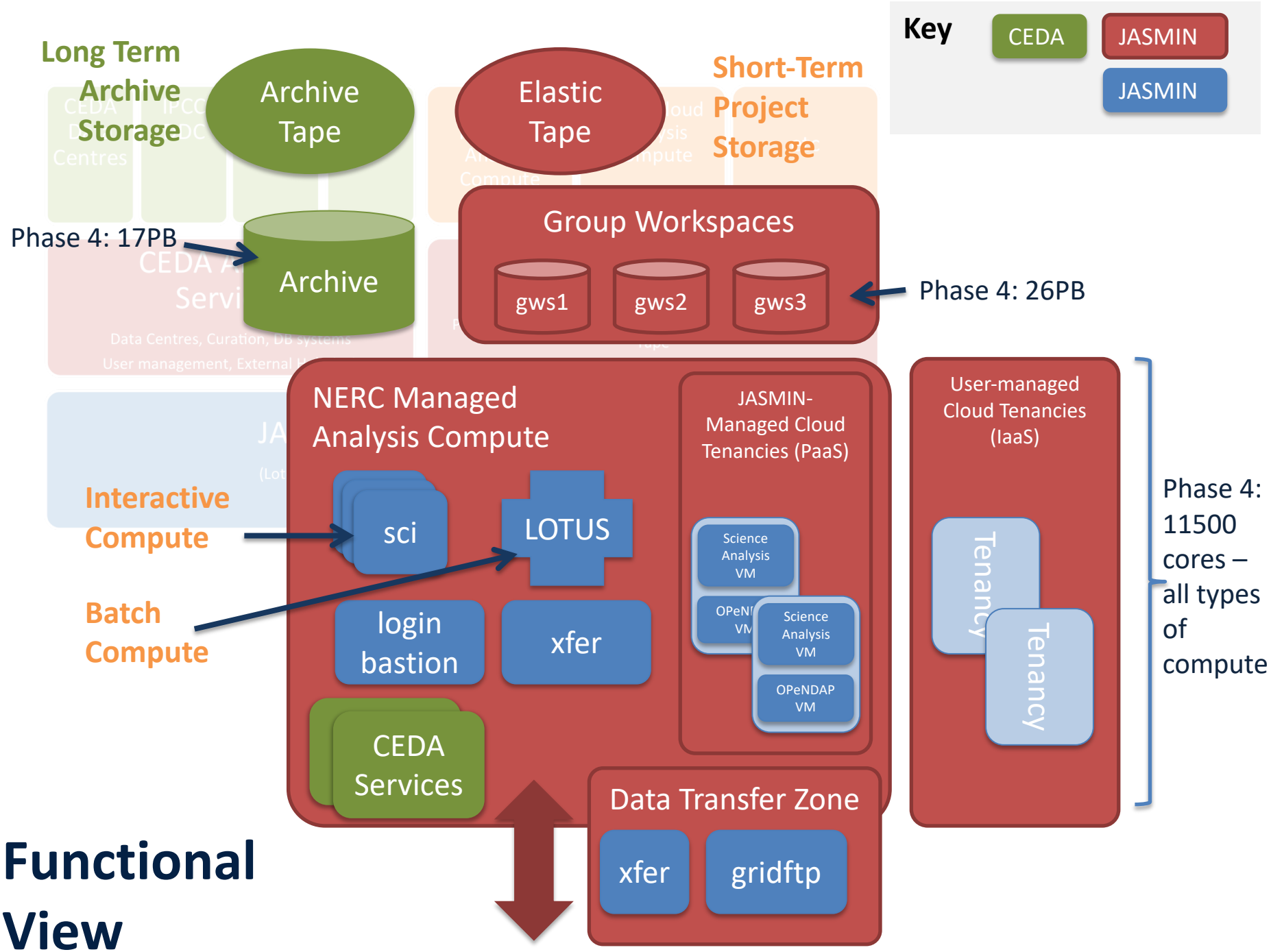


JASMIN Compute and Storage

Managed Compute (Lotus and GWSs, Tape Store + Data Transfer Zone),

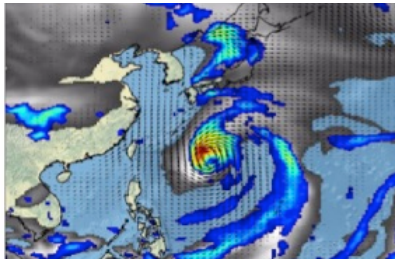
Community Cloud + Internal Helpdesk

NERC SCIENCE OF THE ENVIRONMENT

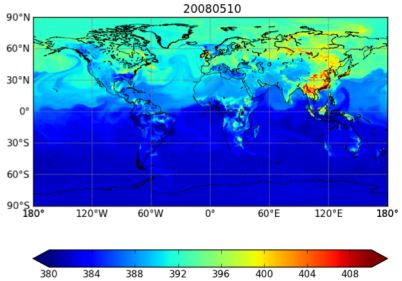


Functional View

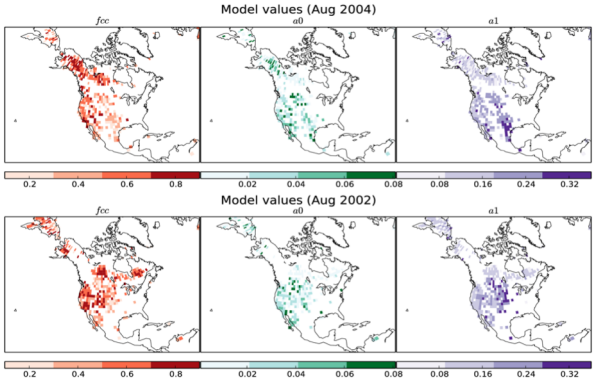
~150 Science projects on JASMIN to date



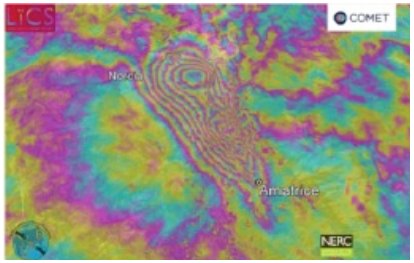
High Res Climate Model analysis



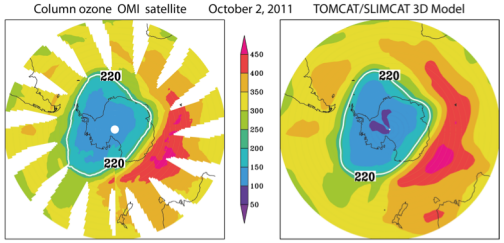
Regional carbon balance on a global scale



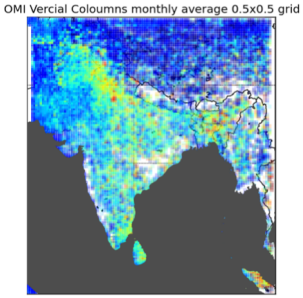
Deriving the impact of fire on vegetation from earth observation data



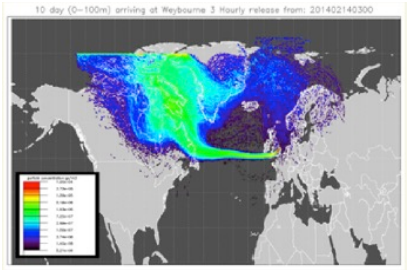
Fault analysis



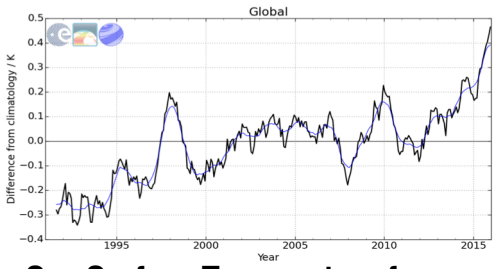
Antarctic Ozone hole: model vs. observations



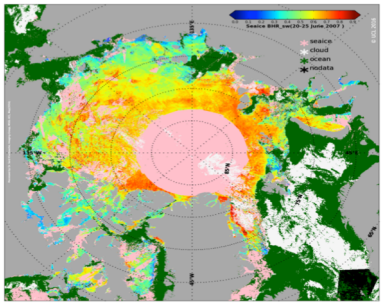
Understanding oxidant chemistry over the Indian subcontinent



Atmospheric dispersion

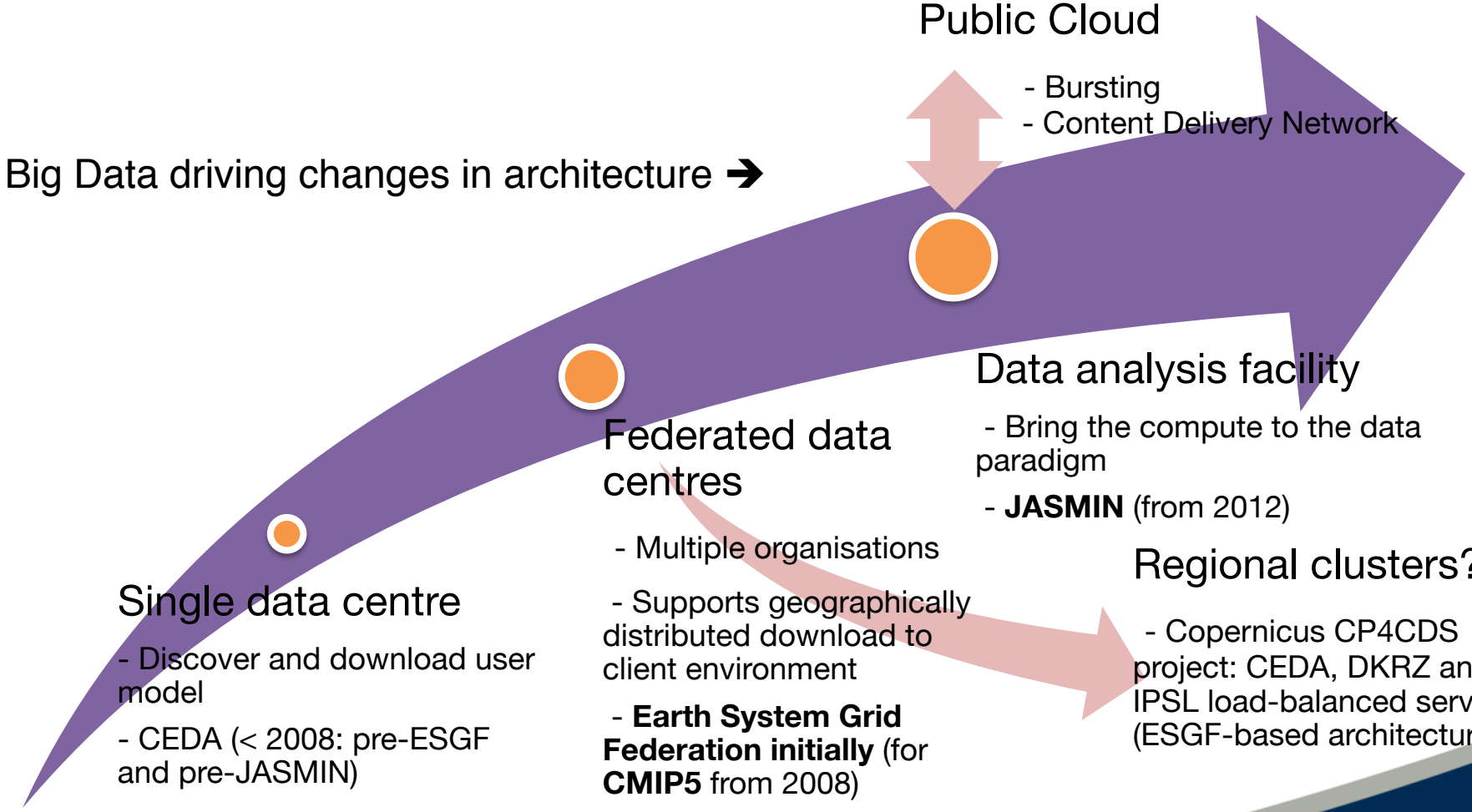


Sea Surface Temperature from satellite observations



Climate variables from European and US instruments/satellites

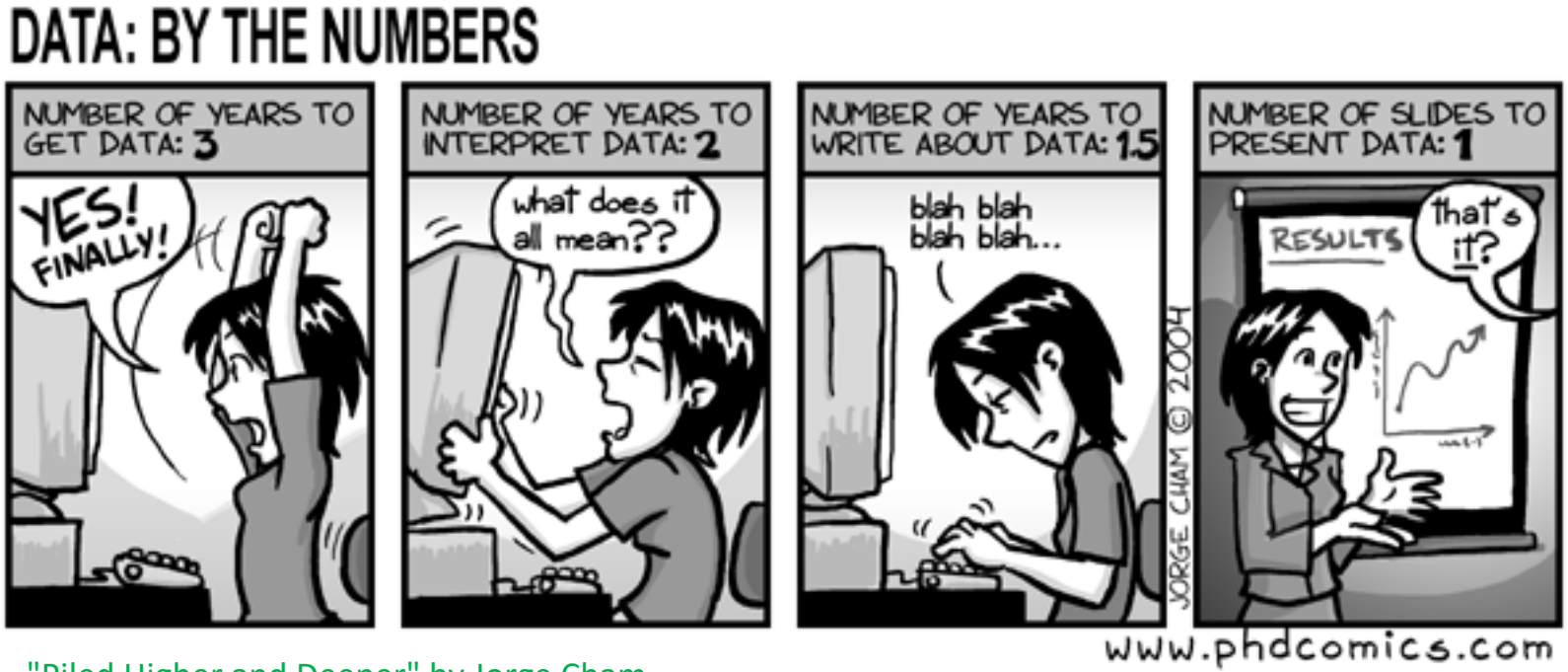
Evolution in models for data distribution and analysis



JASMIN Challenges

- Data intensive computing facility
  - Storage is a major cost
  - Phase 4 introduces more cost effective storage technologies for future scaling: object storage and scale-out file system
- Balancing "bleeding edge" technology with demands of operational service
  - Research infrastructure, not 99.999% available
  - Small team (capital-heavy)
- Diverse user community
  - Wide range of workflow requirements
  - Wide range of user skill level
  - Training, user education, engagement

Creating a dataset is hard work!



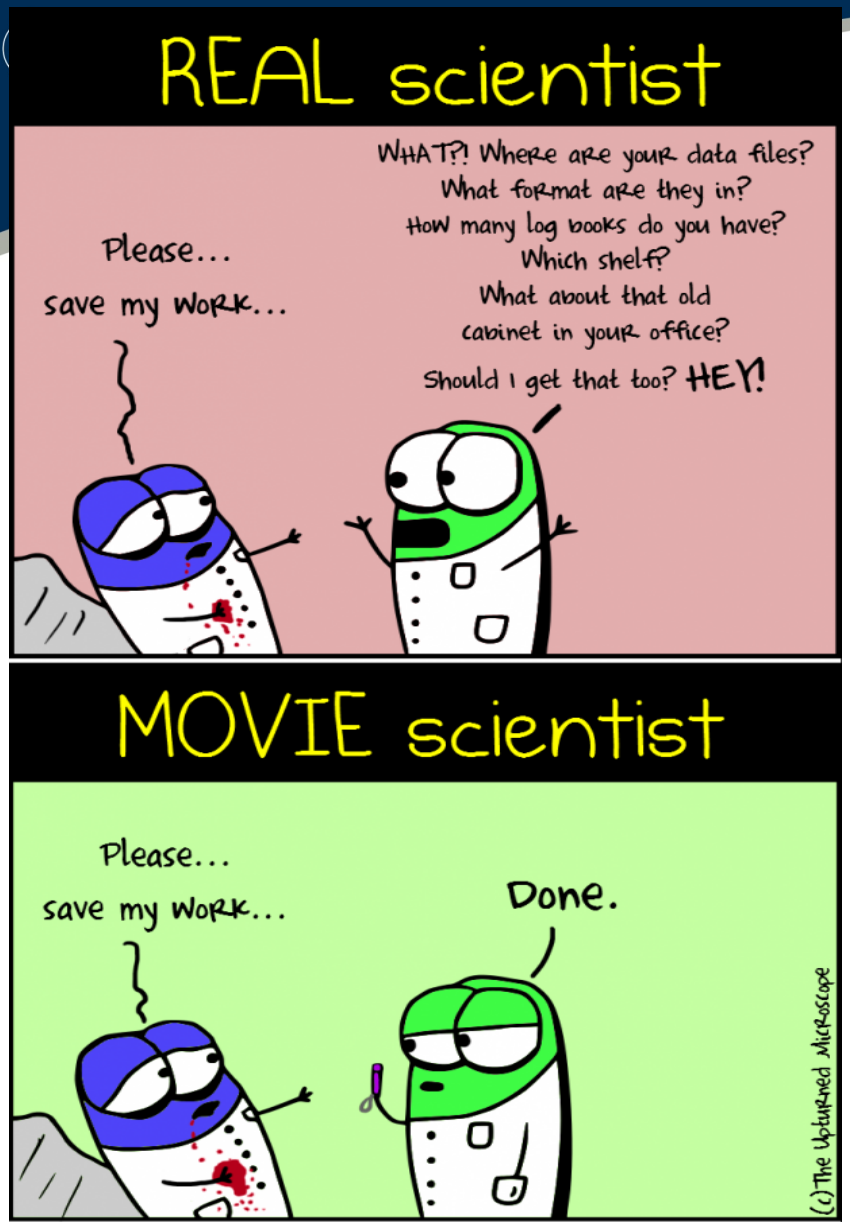
"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

Documenting a dataset so that it is usable and understandable by others is extra work!

Don’t just create a bit bucket

“When required to make the data available by my program manager, my collaborators, and ultimately by law, I will grudgingly do so by placing the raw data on an FTP site, named with UUIDs like 4e283d36-61c4-11df-9a26-edddf420622d. I will under no circumstances make any attempt to provide analysis source code, documentation for formats, or any metadata with the raw data. When requested (and ONLY when requested), I will provide an Excel spreadsheet linking the names to data sets with published results. This spreadsheet will likely be wrong -- but since no one will be able to analyze the data, that won't matter.”

- http://ivory.idyll.org/blog/data-management.html



http://theupturnedmicroscope.com/comic/real-vs-movie-scientist-3/

Work with your users

- Build things in a modular fashion which can be easily reused for different communities
- Start small and build up
- Minimise data friction – make it easy for users to submit/download/use data

- If there's no users – what's the point of the repository?

(some of) What we've learned

- Train your depositors in good data management
- Use off the shelf solutions where possible
- Not all data should be kept (data value checklist)
- It's easier to create accurate metadata at the beginning of the project
- Providing secure workspaces makes researchers *more* likely to share their data
- Researchers need incentives to put their data in a repository (especially if there is extra work needed to do so)
- Standard tools and formats make life easier for everyone

"Really, they' d be rushing around collecting revisions to go into the next scheduled quarterly public data update, not publishing them immediately, but you have to embellish things a little for Hollywood."

https://xkcd.com/2029/

Thanks!

Any Questions?

sarah.callaghan@stfc.ac.uk
@sorcha_ni