


JASMIN (STFC/Stephen Kill)

Software Development at the Centre for Environmental Data Analysis

RAL Site Software Engineering Community Meeting

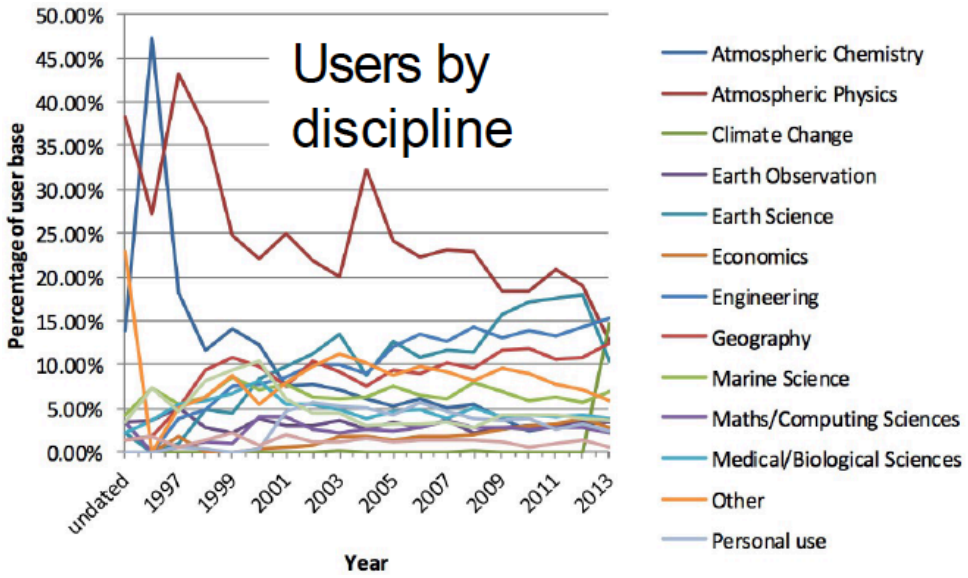
12 February 2018

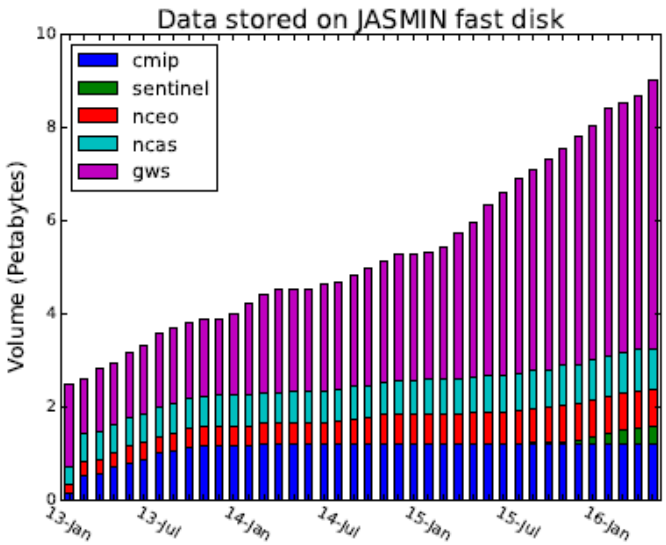
Philip Kershaw (on behalf of CEDA)

NCAS/NCEO, Centre for Environmental Data Analysis, RAL Space

[Thanks and credit to STFC Scientific Computing Department who deploy and operate the JASMIN infrastructure on behalf of CEDA]

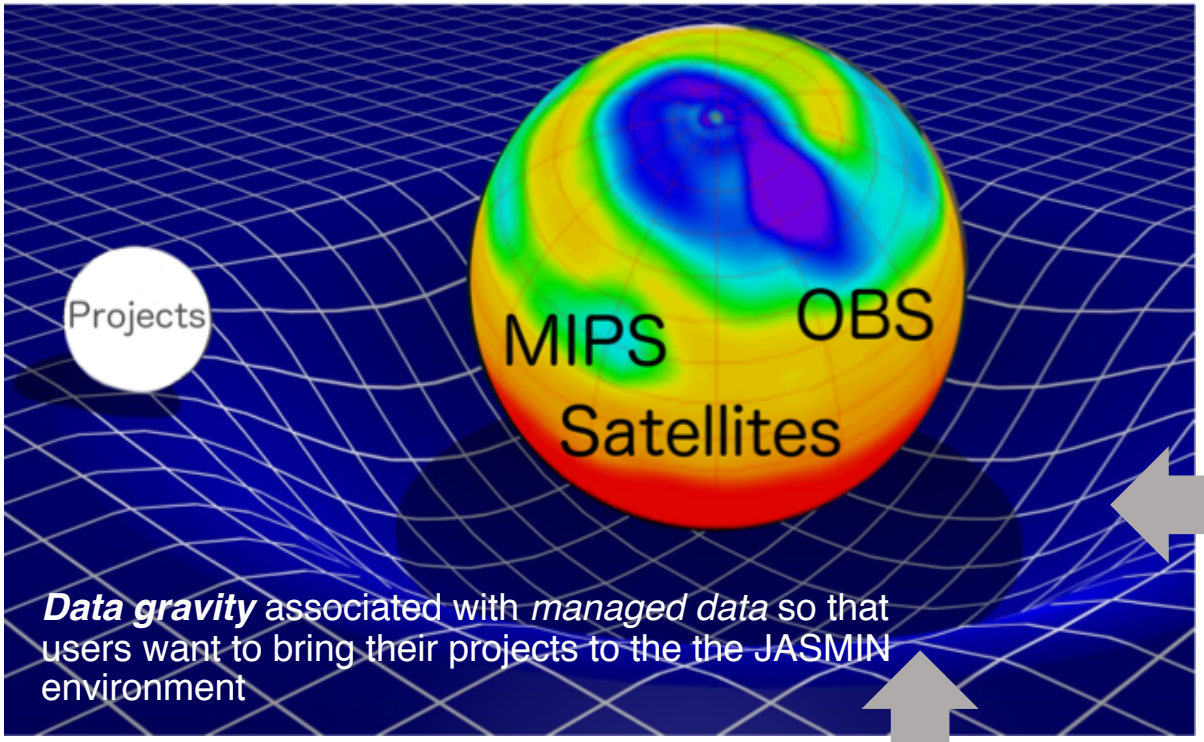
Data Growth and Diversification of user community





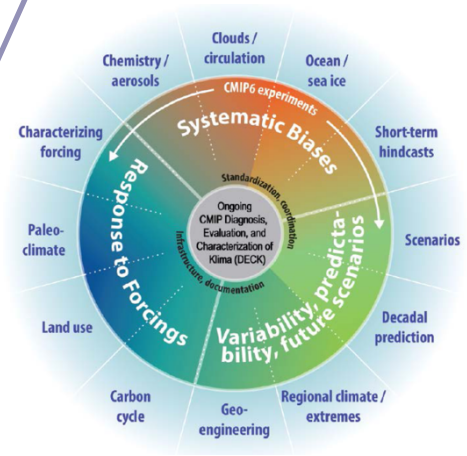
2013-2016 increasing data storage on JASMIN, in Group Workspaces (GWS) and archive

JASMIN as a Data Commons



CEDA CREPP system to ingest from MetOffice Hadley Centre

CMIP6



Climate Models

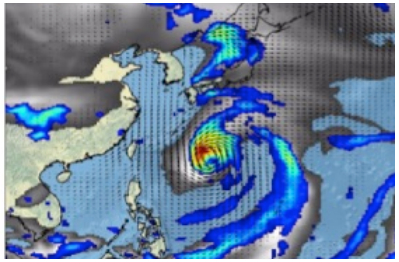
Couple-Model Intercomparison Projects

Sentinel Earth

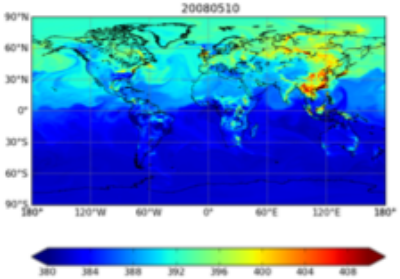
Observation Data

Sentinel missions data rate: ~6PB/year

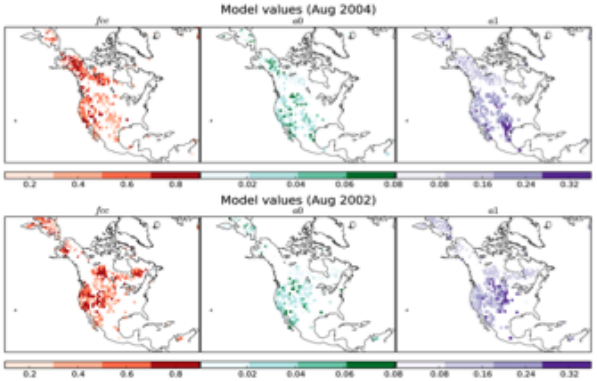
~150 Science projects on JASMIN to date



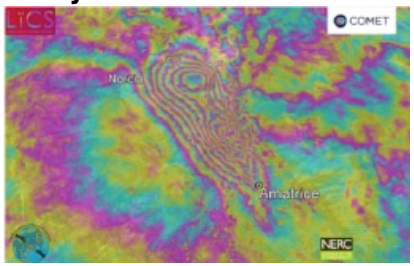
High Res Climate Model analysis



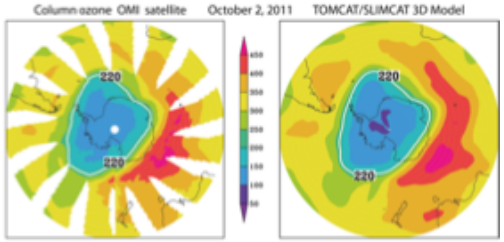
Regional carbon balance on a global scale



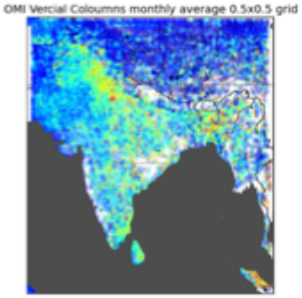
Deriving the impact of fire on vegetation from earth observation data



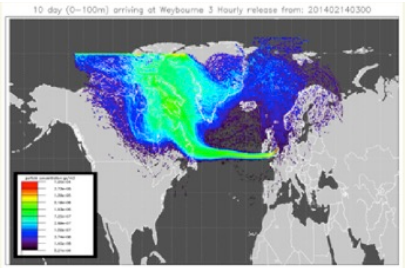
Fault analysis



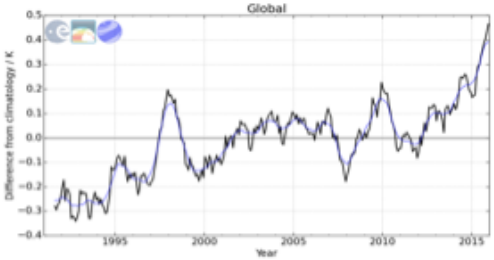
Antarctic Ozone hole: model vs. observations



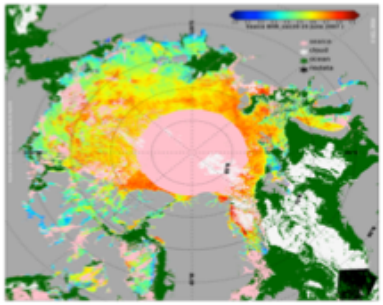
Understanding oxidant chemistry over the Indian subcontinent



Atmospheric dispersion



Sea Surface Temperature from satellite observations



Climate variables from European and US instruments/satellites

JASMIN usage: Cloud

ESA Forestry Thematic Exploitation Platform

ESA Climate Change Initiative Open Data Portal

Majic interface to Jules Land-surface model on JASMIN

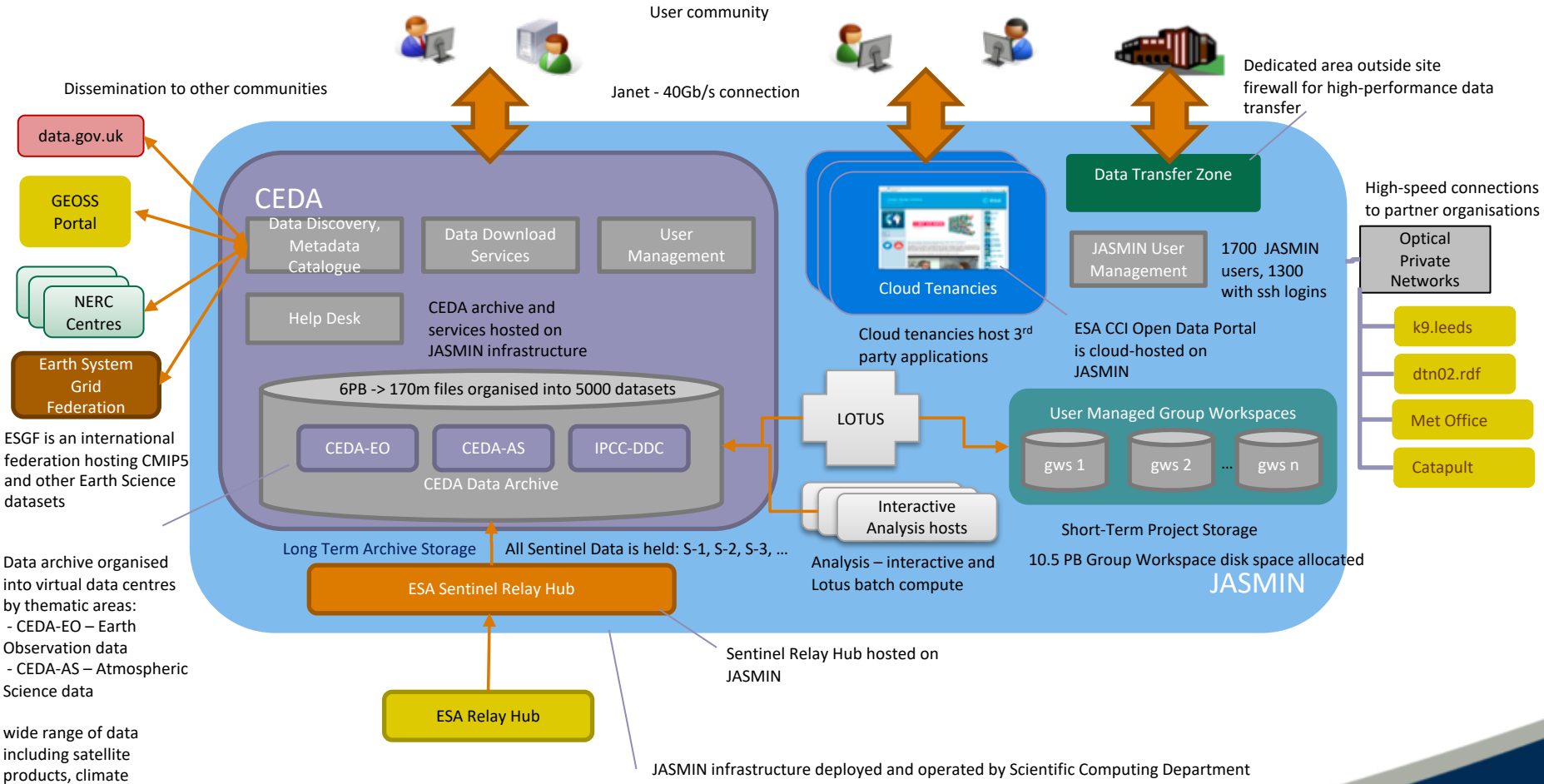
ESA Polar Thematic Exploitation Platform

Attendees at ESA Summer school, ESRIN used OPTIRAD Jupyter Notebook environment

– Credit ESA

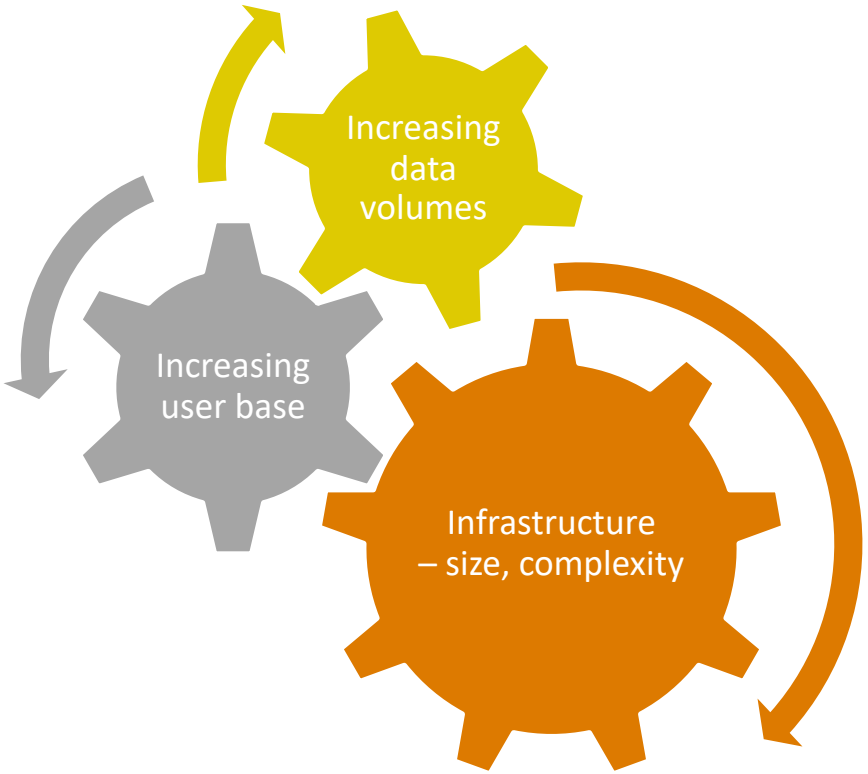
EOS Cloud – Desktop-as-a-Service for Environmental Genomics

JASMIN and CEDA



Challenges and new developments to address them (1)

Key technologies to address specific challenges

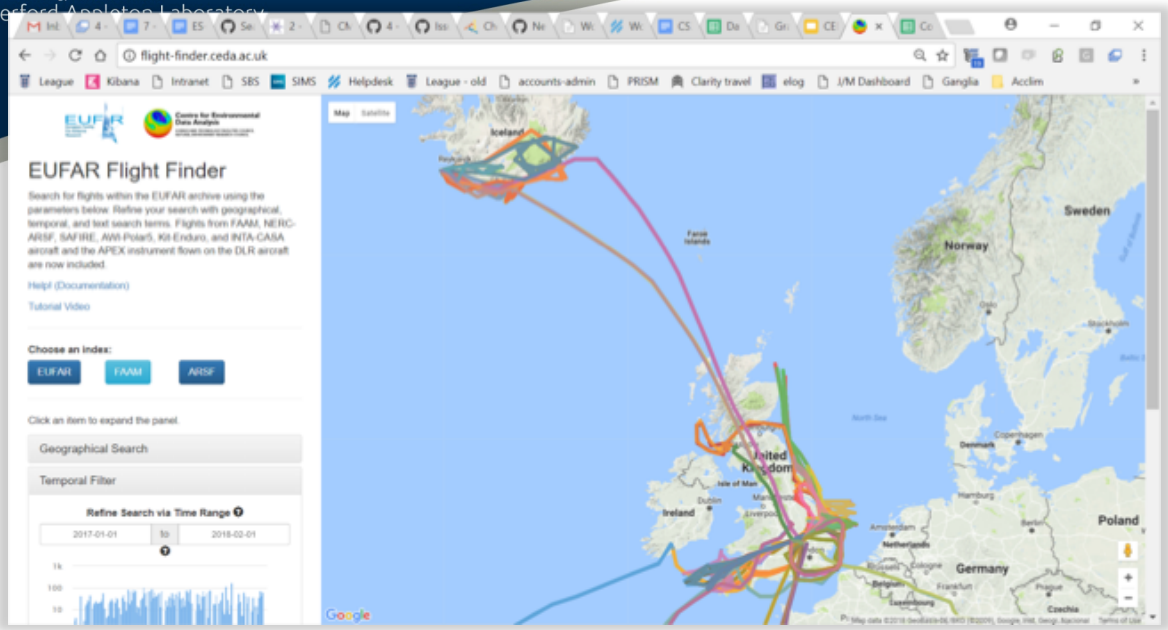


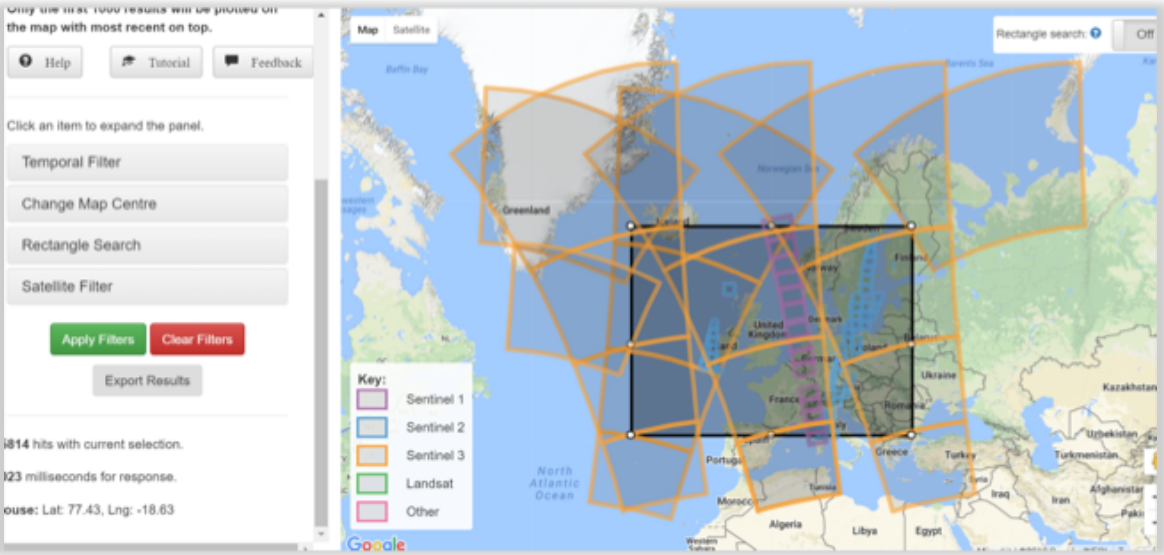
Data discovery and Cataloging

Object Storage

Containers and Container orchestration

On-prem ⇳ Public Cloud portability

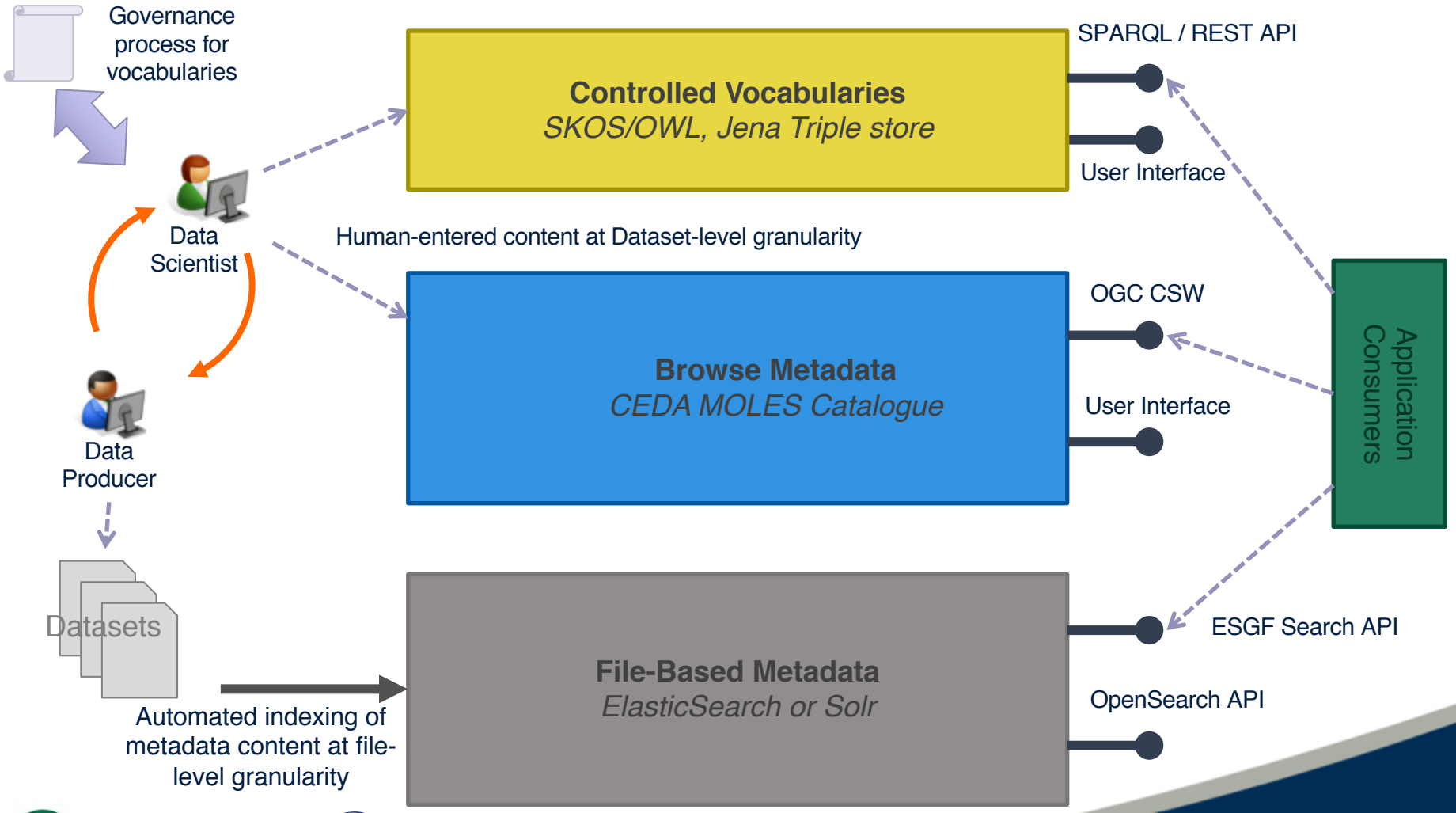

flight-finder.ceda.ac.uk
EUFAR Flight Finder
Search for flights within the EUFAR archive using the parameters below. Refine your search with geographical, temporal, and text search terms. Flights from FAAM, NERC-ARSF, SAFIRE, AWI-Polar5, Kit-Enduro, and INTA-CASA aircraft and the APEX instrument flown on the DLR aircraft are now included.
Help! (Documentation)
Tutorial Video
Choose an index:
EUFAR
FAAM
ARSF
Click an item to expand the panel.
Geographical Search
Temporal Filter
Refine Search via Time Range
2017-01-01
to
2018-02-01
Iceland
Norway
Sweden
Ireland
United Kingdom
Netherlands
Germany
Poland
Denmark
Google


the map with most recent on top.
Help
Tutorial
Feedback
Click an item to expand the panel.
Temporal Filter
Change Map Centre
Rectangle Search
Satellite Filter
Apply Filters
Clear Filters
Export Results
814 hits with current selection.
23 milliseconds for response.
ouse: Lat: 77.43, Lng: -18.63
Map
Satellite
Rectangle search:
Key:
Sentinel 1
Sentinel 2
Sentinel 3
Landsat
Other
Greenland
United Kingdom
Ukraine
Kazakhstan
Uzbekistan
Turkmenistan
Afghanistan
Iran
Iraq
Turkey
Greece
Algeria
Libya
Egypt
Google

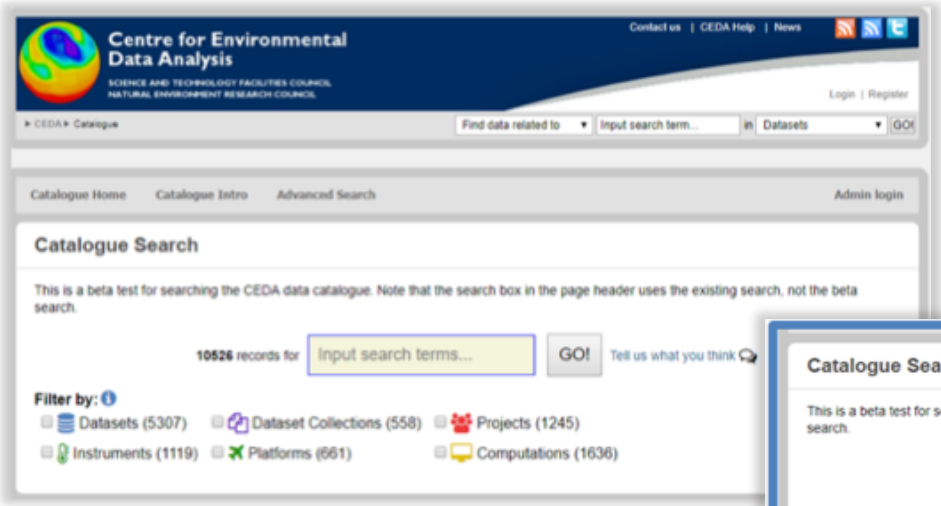

EUFAR Flight Finder &
CEDA Satellite Data Finder
Wendy Garland, Richard Smith and Ag Stephens
NCAS/NCEO, Centre for Environmental Data Analysis, RAL Space, STFC, UK
Introduction
How the EFF and Satellite Finder works
Challenges
Conclusions and future plans
http://flight-finder.ceda.ac.uk
http://geo-search.ceda.ac.uk
Contact: Wendy Garland
wendy.garland@ncas.ac.uk

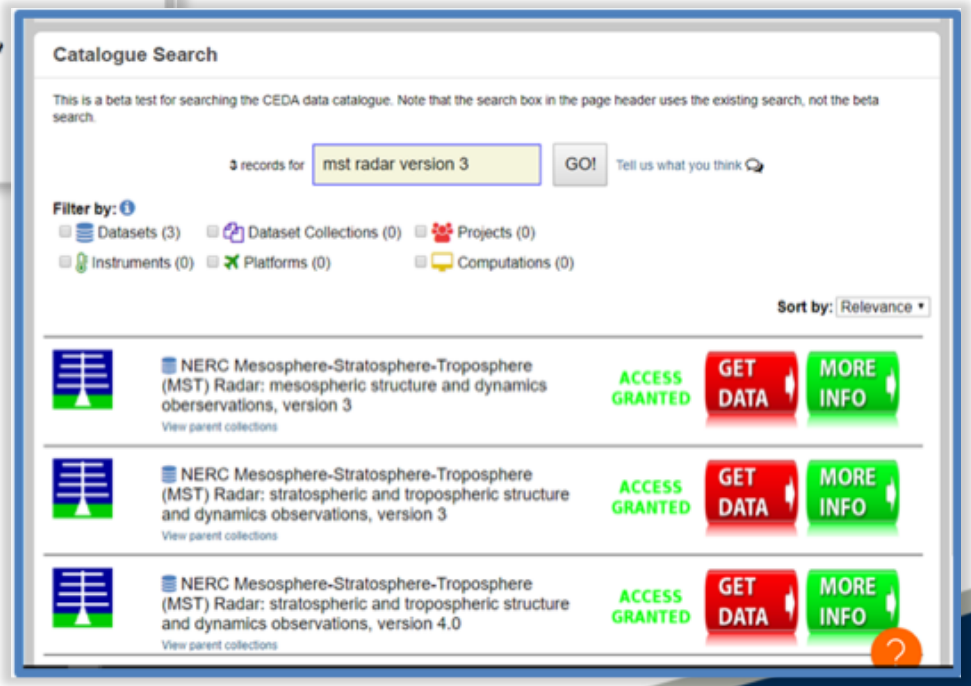
elastic

Sourcing information for Data discovery



Catalogue search



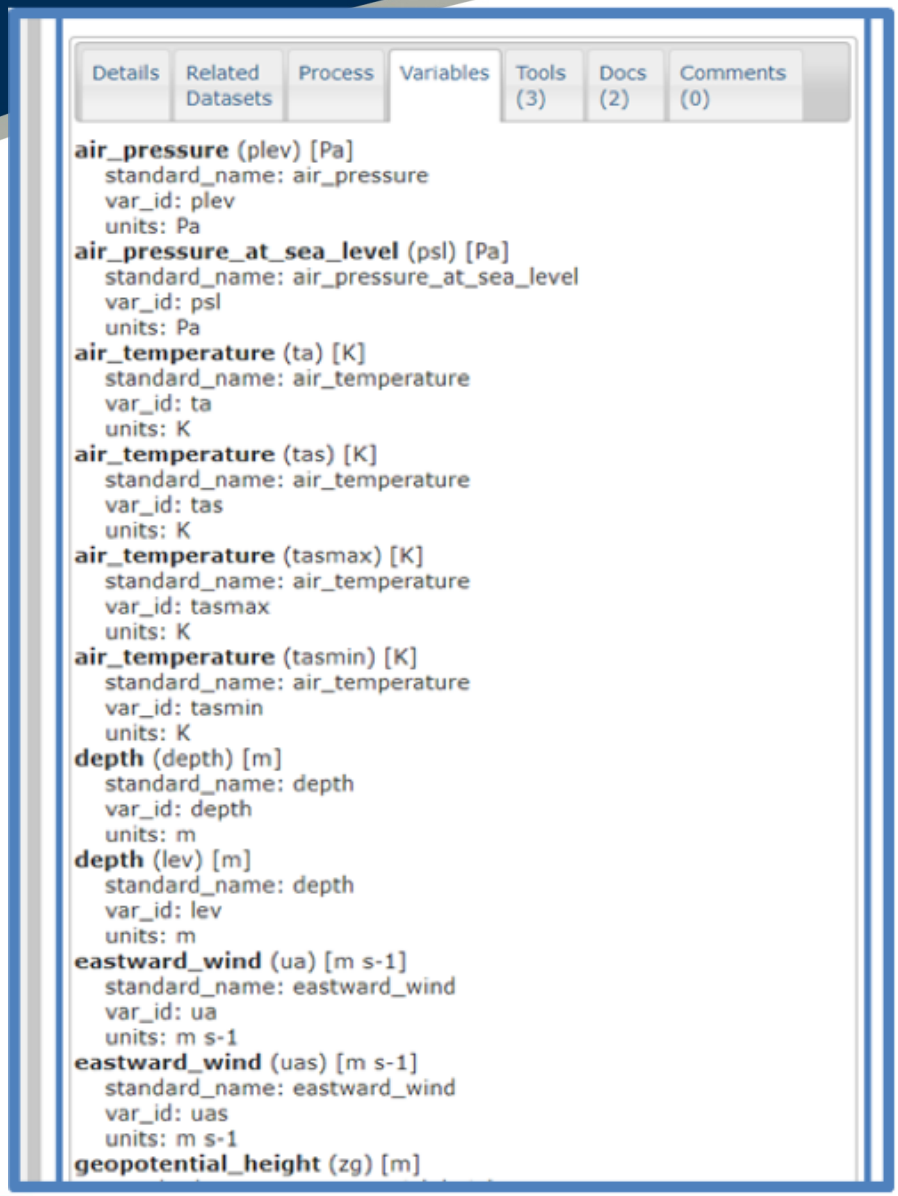


python™

&

django

Dataset and File-level metadata ingration



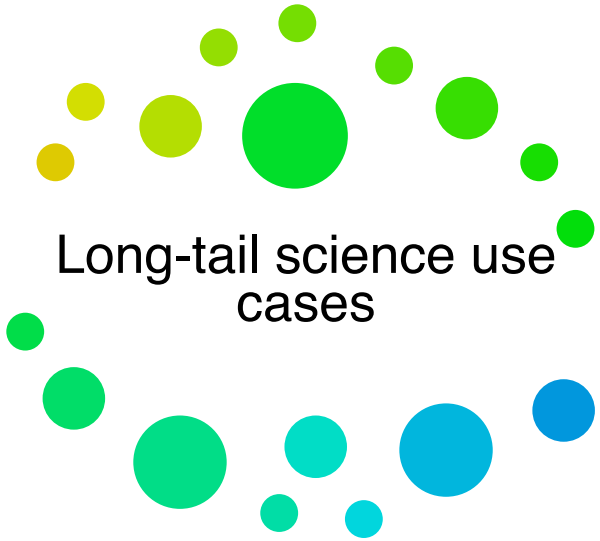
python

&

django

elastic

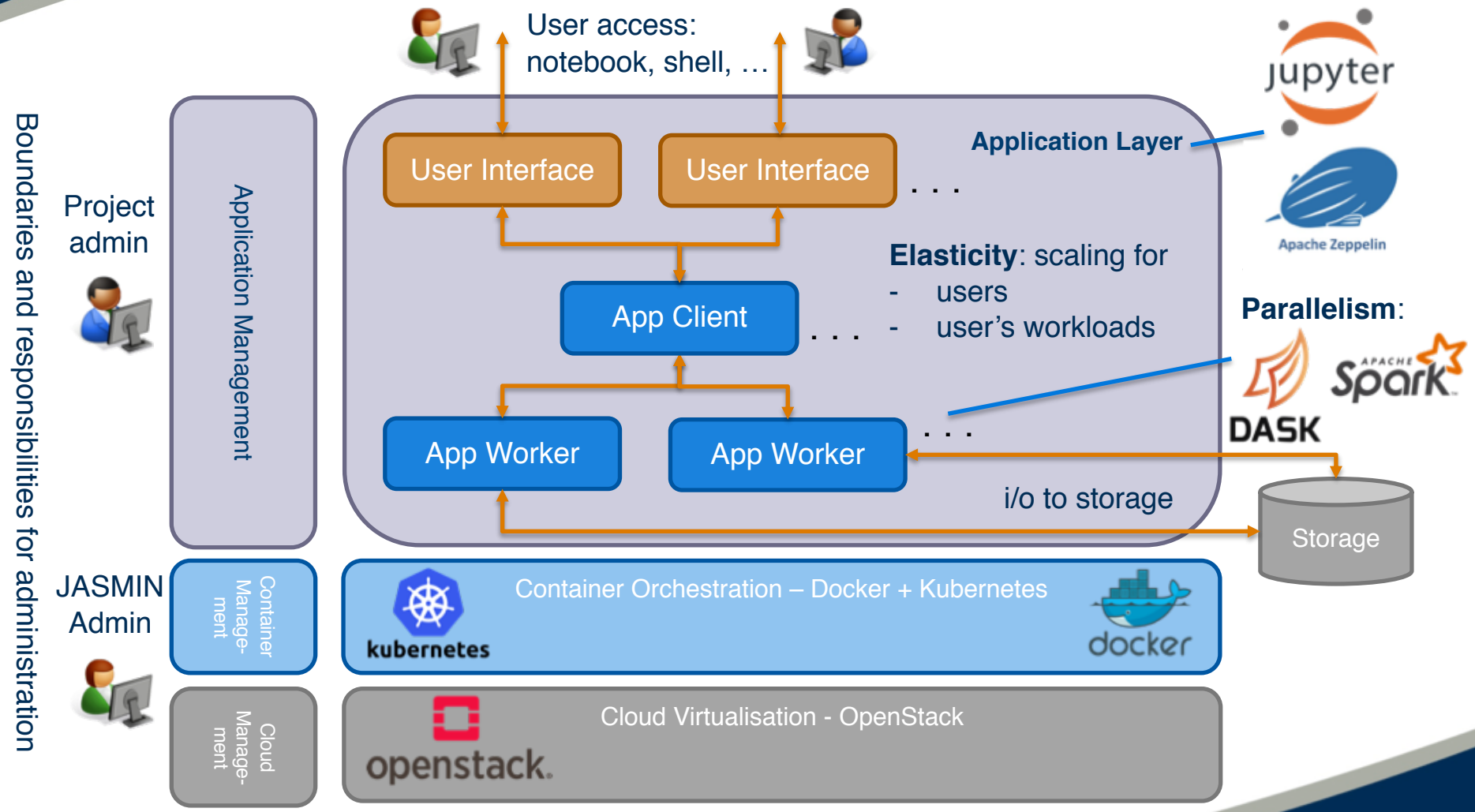
Challenges and new developments to address them (2)



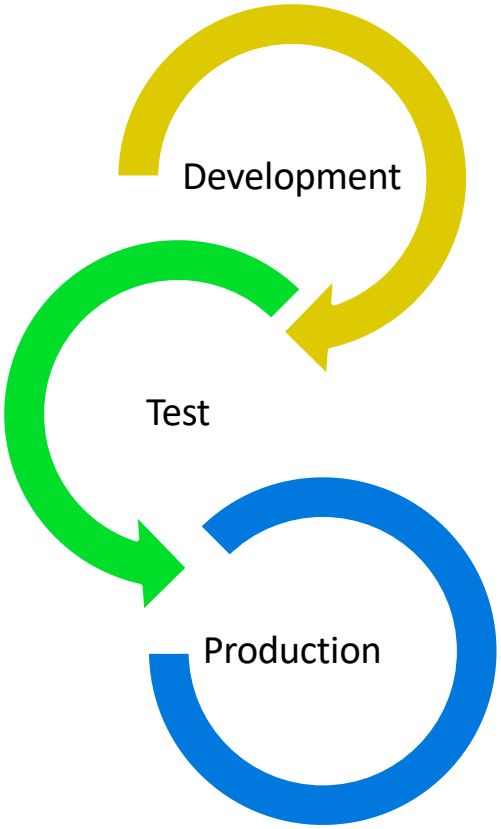
- Need for effective exploitation of parallelism to deliver demonstrable benefit over user's desktop/laptop
- Need for ease of use
- Intuitive user interfaces

- Software-as-a-Service model
- Desktop app-style user experience
- dynamic provisioning of clusters
- Parallel programming libraries

Data Analytics



Development challenge: size and complexity of code and systems footprint



- CEDA Development team
  - 15 people includes data scientists, ops and managers – approx. 7 f/t s/w development
- Languages
  - Python predominates, all new projects Python 3
  - One Cython project
  - JS - React
  - Some Java
- Development environments: PyCharm, PyDev (Eclipse), CLI + editors
- All projects Open Source by default
  - https://github.com/cedadev/
  - Private git for deployment-sensitive content (e.g. Ansible playbooks)
- Training in the user community:
  - Introduction to Scientific Computing Course
- Build, test, integrate, operate
  - Vagrant + Ansible
  - Cloud dev tenancy
  - Standardised on RedHat 6/7
  - Planned: Docker + Kubernetes (OpenShift)
  - Production checkout process and documentation
  - Integrating code tests into Icinga/Nagios operational monitoring

Summary

- JASMIN: data gravity, a data commons for environmental sciences
- Challenges with respect to running at scale:
  - Data volumes
  - Numbers of users
  - Generation and indexing of content for effective discovery and understanding of data for users
  - Effective use of parallelism for long-tail of science users
  - Increasing footprint of code and systems to manage for development and operations
- New infrastructure services to address challenges:
  - Evolution of data discovery and cataloguing systems, AI exploitation?
  - Virtual Research Environments
  - Object store migration
  - Development and operations: Increasing Automation – virtualisation containers and container orchestration

Great CEDA Data Discovery Hackathon (coming soon)

Further Information

CEDA team

- CEDA and JASMIN:
  - http://www.jasmin.ac.uk/
  - http://www.ceda.ac.uk/
- Github:
  - https://github.com/cedadev/
- JASMIN paper

  Lawrence, B.N. , V.L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard, and A. Stephens. **Storing and manipulating environmental big data with JASMIN.** *Proceedings of IEEE Big Data 2013, p68-75,* doi:10.1109/BigData.2013.6691556
- philip.kershaw@stfc.ac.uk, @PhilipJKershaw