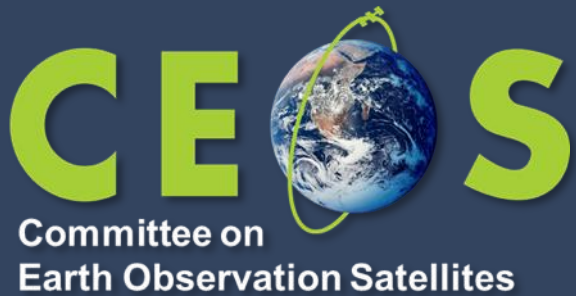


# CAL: The CEOS Analytics Lab



# CEOS Analytics Lab (CAL)



Services ▾

## CEOS Analytics Lab

Empowering exploration and scalable analysis of Earth observation data

The CEOS Analytics Lab is a multiuser gateway for spatial data science powered by EASI. Every user is provided a customized JupyterLab environment to easily load EO data products and seamlessly scale to additional computational nodes through the Dask Gateway.

Login [↗](#)

### Support

We are here to help the CEOS community succeed. If you have a potential application or need a platform for EO data analysis we would like to hear from you. Please be aware that the CEOS Analytics Lab is still undergoing changes and we are interested in gathering your feedback.

- Currently GPU enabled machine learning instances are not available by default. These can easily be enabled by request. Please get in touch if your analysis requires machine learning.
- Support Requests: If you need to submit a support request you can find the form in the Services menu at the top right of the page.
- Account Creation: Accounts can be created by filing a service request to access our platform and its features.
- Collaboration with CEOS Working Groups: We are open to collaborating with CEOS working groups to address specific requirements and ensure seamless integration.
- Training Opportunities: Take advantage of our training sessions designed to help you make the most of our platform and its capabilities.
- Scalability Options: Requests for larger instances will be accommodated to meet your evolving needs. Submit a service request for additional resources if you are reaching the limits of the provided options.

Special thanks to our partners who made the CEOS Analytics Lab possible



CAL is powered by CSIRO's EASI technology

This is a research platform provided on a best effort level. Although you may store your data here it is your responsibility to ensure it is backed up.

- ❖ Initiated in April 2020 as a CEOS WGISS initiative, CAL is a data and analytics platform. It offers access to the Open Data Cube, a hosted JupyterHub environment, Dask scaling, and customized ARD pipelines all running on the AWS cloud. Both CPU and GPU processing is available.
- ❖ There are currently ~75 registered users!

# Purpose



- ❖ CAL is a platform to serve the CEOS community and facilitate collaborative Earth Observation analysis
- ❖ Objectives:
  - Provide an open Jupyter notebook environment
  - Simplify loading of Earth observation data
  - Minimize analysis setup time and overhead

The screenshot displays a Jupyter Notebook environment. The top part of the notebook shows the title "CEOS Analytics Lab" and the subtitle "Cloud Coverage Statistics". Below the title, there is a code cell with the following Python code:

```
[8]: import matplotlib.pyplot as plt
draw_ditter_plot(clean_percent, dates, color_map = "BuPu", figsize = (20,1))
draw_ditter_plot(clean_percent, dates, color_map = "BuPu", sigma = 1, figsize = (20,1))
draw_ditter_plot(clean_percent, dates, color_map = "BuPu", sigma = 5, figsize = (20,1))
draw_ditter_plot(clean_percent, dates, color_map = "BuPu", sigma = 10, figsize = (20,1))
```

Below the code cell, there are four vertically stacked bar charts. Each chart shows "Cloud Coverage" on the y-axis (ranging from 0.0 to 1.0) and "Year" on the x-axis (ranging from 2014 to 2024). The charts display a dense pattern of vertical bars, with the top chart showing the highest coverage and the bottom chart showing the lowest coverage. The bars are colored in a gradient from blue to purple.

At the bottom of the notebook, there is another code cell with the following Python code:

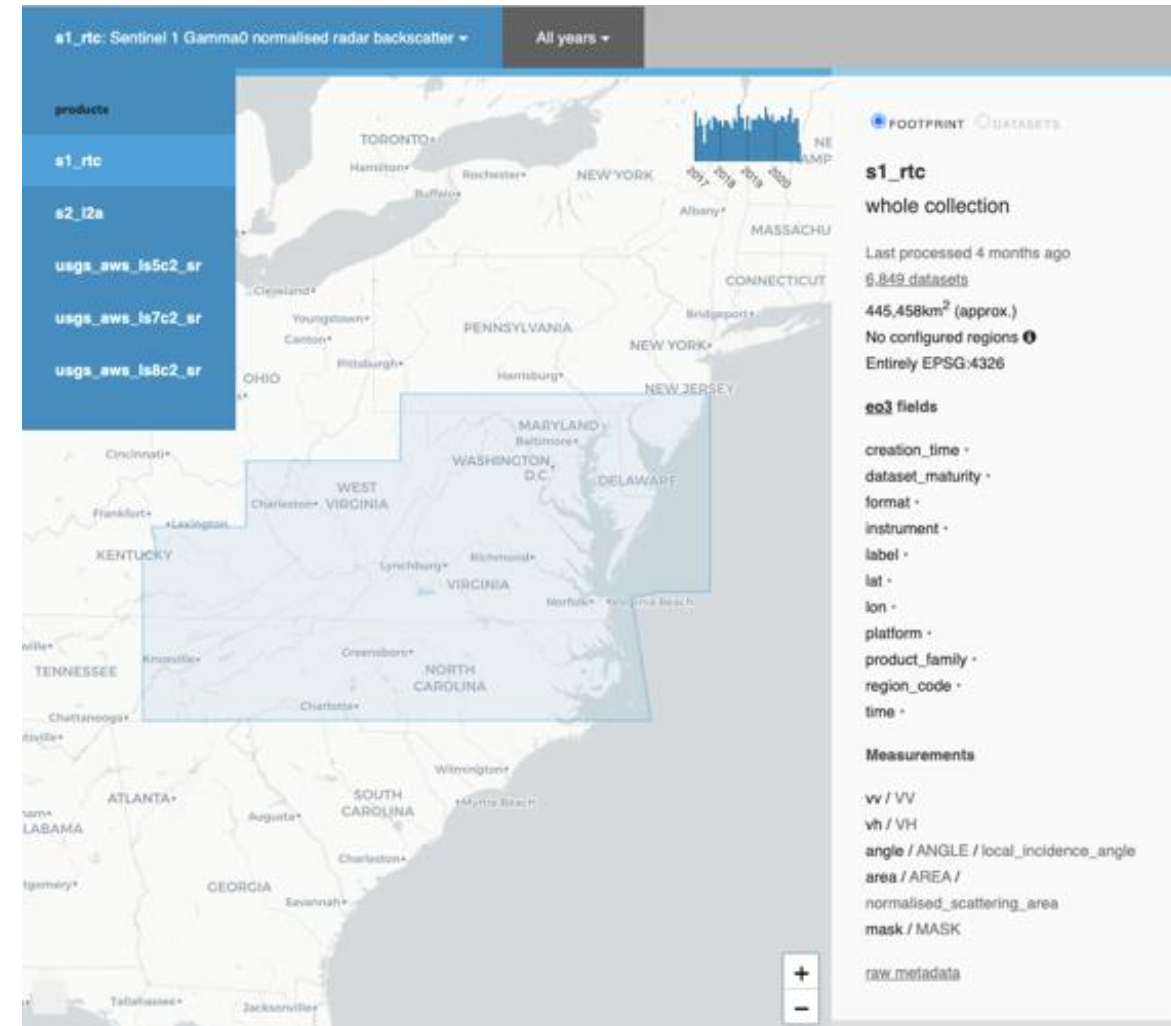
```
[15]: import os
import sys
sys.path.append(os.path.expanduser("../scripts"))

import seaborn as sns

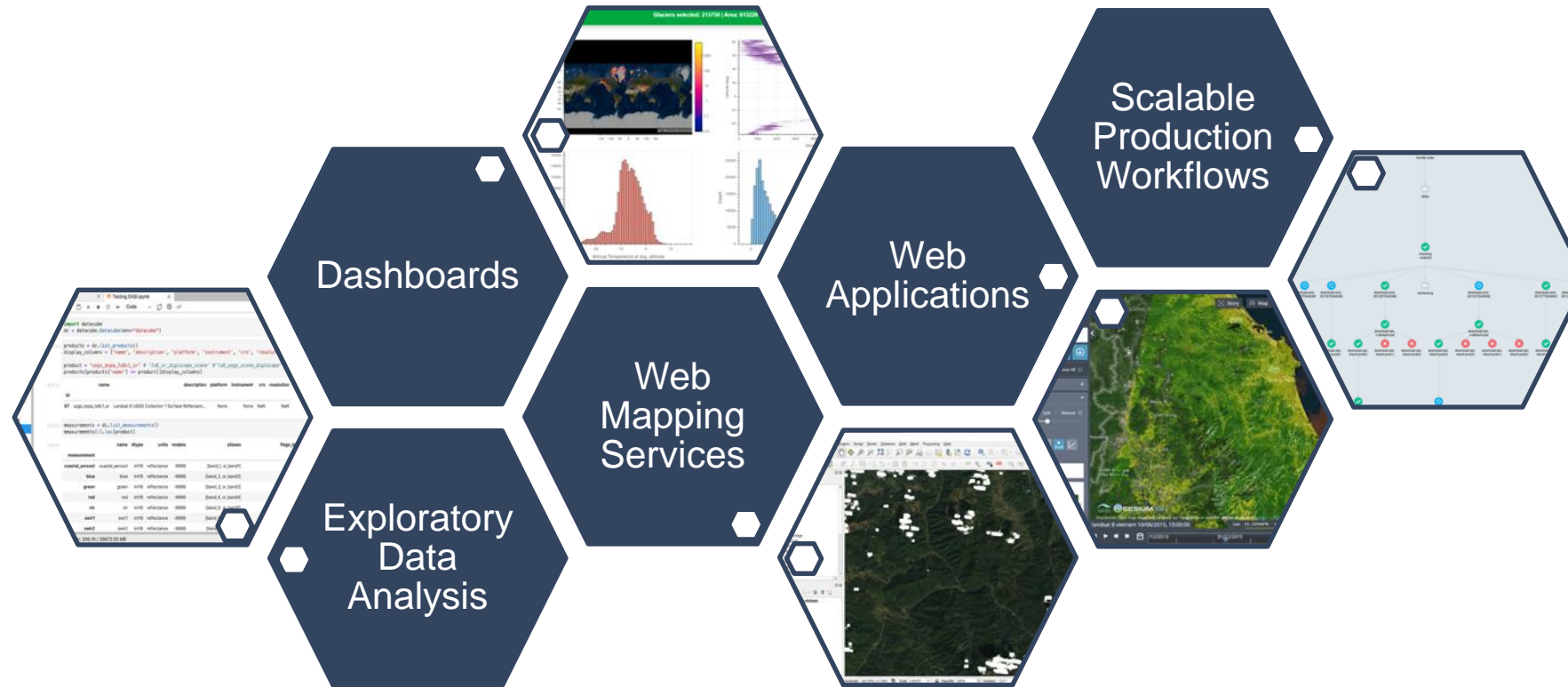
os.environ["USE_PYGEOS"] = "0"
from datacube.utils.aws import configure_s3_access
```



- ❖ **CAL** currently supports the **COAST** (Chesapeake Bay study) project for CEOS and working with the CEOS **Ecosystem Extent Pilot Project**
- ❖ Potential future users include: **WGCV** (DEMIX Cal-Val campaigns), **DE-Americas** (Caribbean Pilot project), and CEOS **Ecosystem Extent Pilot Project**
- ❖ **Datasets** currently available: Landsat, Sentinel-2, MODIS, Sentinel-3, Sentinel-1 (CARD4L with RTC), Copernicus DEM, and NASA DEM



- ❖ CAL is built using the Open Data Cube software and CSIRO's Earth Analytics, Science and Innovation platform



Powered by Open Data Cube and the Python data science ecosystem

# New & potential analytics capabilities



- ❖ GPU processing with AWS GPU nodes
- ❖ Scalable across multiple computation nodes
- ❖ Additional scientific programming options with R
- ❖ New machine learning capabilities
- ❖ Processing Pipelines based on Argo Workflows



# Data Science and GIS are Merging



- ❖ Data science tools are being used to tackle big EO data
- ❖ Notebooks are a ubiquitous data science tool
- ❖ Jupyter notebooks are web based interactive development environment where textual explanations, graphical outputs and code are presented together
- ❖ Enables users to conduct analyses with common Python data science tools

Will GIS eventually merge with 'datascience'?

Discussion

For a long time GIS was a specialization of its own, with distinct tools, datasets, algorithms etc. Will it remain so or will it become more integrated as a branch of general purpose datascience (echoing how postgres is an extension of postgres)

1.3k votes

399 No. Its tools and concepts are too specialized and they will always evolve separately fr...

138 Yes, but it will be decades before it happens

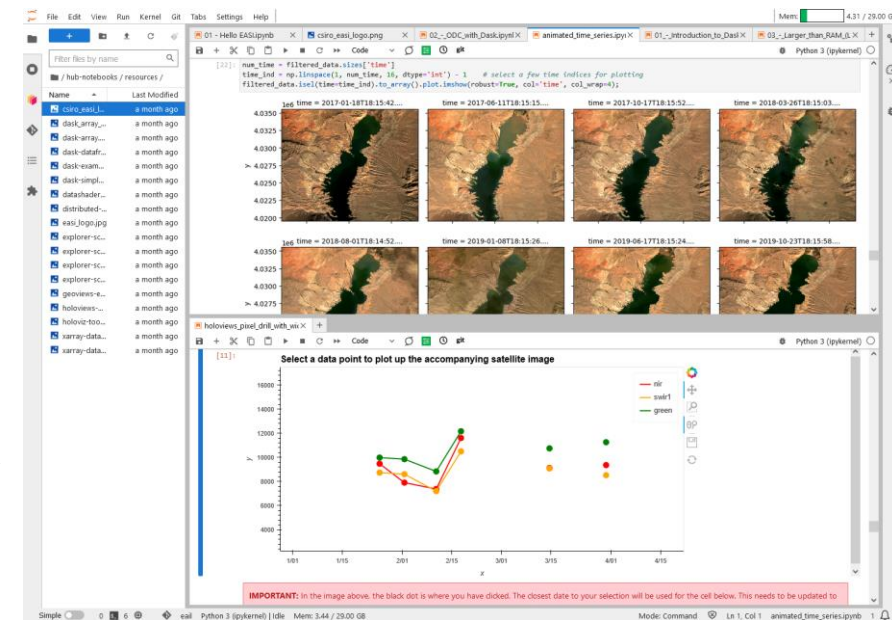
804 Yes, it is happening already ✓

Voting closed 3 months ago

# Analysis Environment: JupyterHub



- ❖ A JupyterHub instance serves notebook environments to multiple users simultaneously
- ❖ User environments are preconfigured with an identical set of libraries
- ❖ The environments are containerized and isolated from other users
- ❖ Simplifies data loading and access when configured with tools such as the Open Data Cube





# Notebook Advantages



- ❖ Allows researchers to ‘play’ with data:
  - It is often important to understand the shape, quality, and type of the raw data when beginning an analysis
  - It is equally useful to be able to explore and understand the effects of intermediate transformations conducted during an analysis
- ❖ The notebook format enables a literate programming paradigm
- ❖ A major advantage of notebooks is the ability to mix plain language explanations and rich content with code that can be executed in place
- ❖ Sharing and executable code promotes reproducibility of scientific results

- ❖ Notebooks contain generic python code that any python interpreter can execute
- ❖ There is a large community of resources and tools for python and Jupyter notebooks
- ❖ Notebooks are a convenient mechanism for transferring and sharing code and programming stories
  - The notebook format is a paper plus everything you need to reproduce a result or modify for you own research



- ❖ Software and hosting interoperability remain a challenge:
  - Data can be named differently or configured
  - Software dependencies may be different – i.e. different hosts may use different tools/versions
  - Much more solvable when using common and open tooling
- ❖ Community standardization is only encouraged:
  - Freedom means customization but also potential differences
  - Groups interested in true interoperability should collaborate with other groups in their field to unify naming conventions, software packages, etc.
- ❖ Dependency management in the python ecosystem is still improving with better tooling

- ❖ CAL is operated by CSIRO Chile with significant support from the Chilean Data Observatory, a public-private-academic partnership founded by the Chilean Government (Ministry of Science, Ministry of Economy), Adolfo Ibáñez University and AWS



Chile

