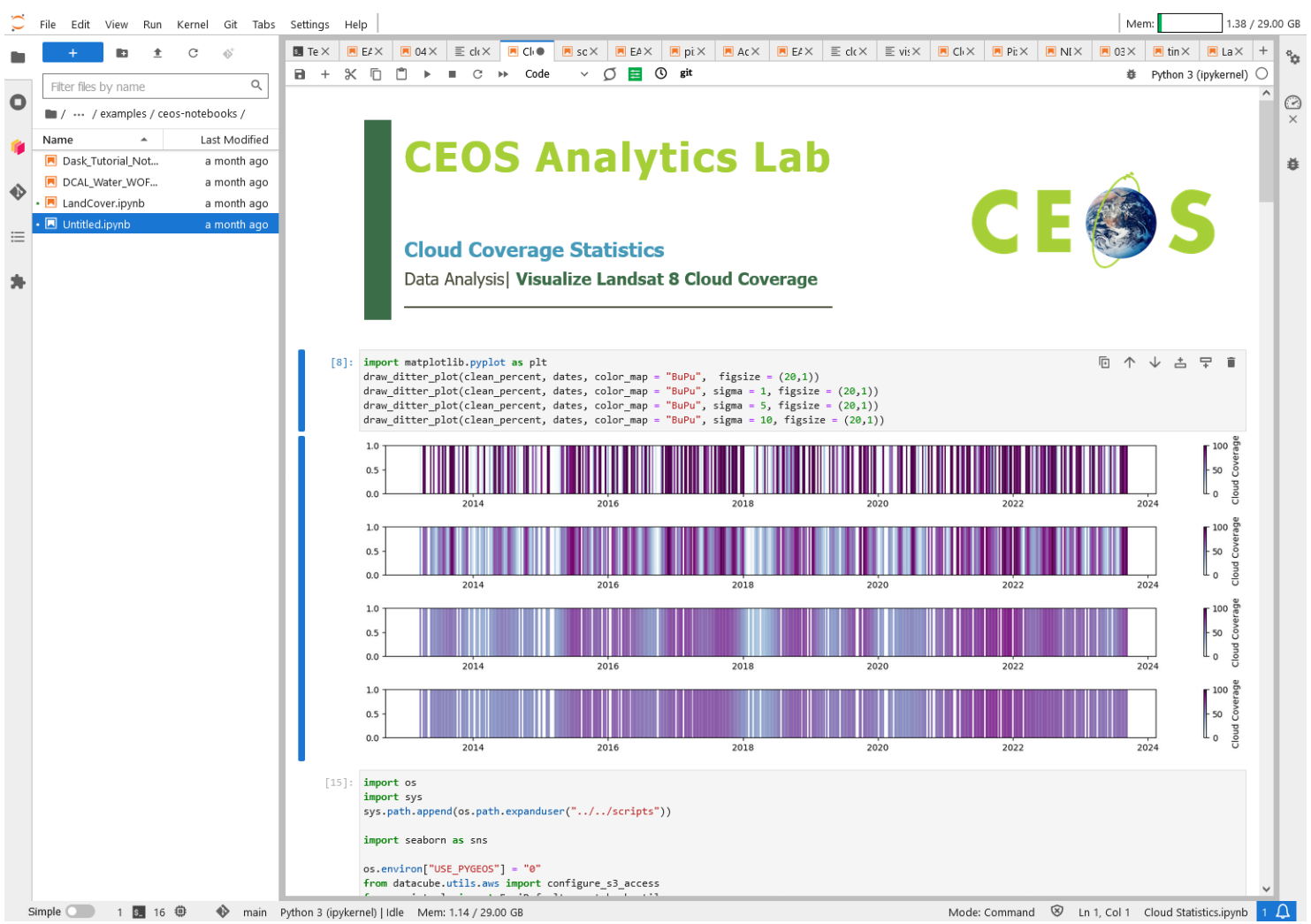
CAL: The CEOS Analytics Lab

CEOS
Committee on
Earth Observation Satellites

CEOS Analytics Lab (CAL)

- 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



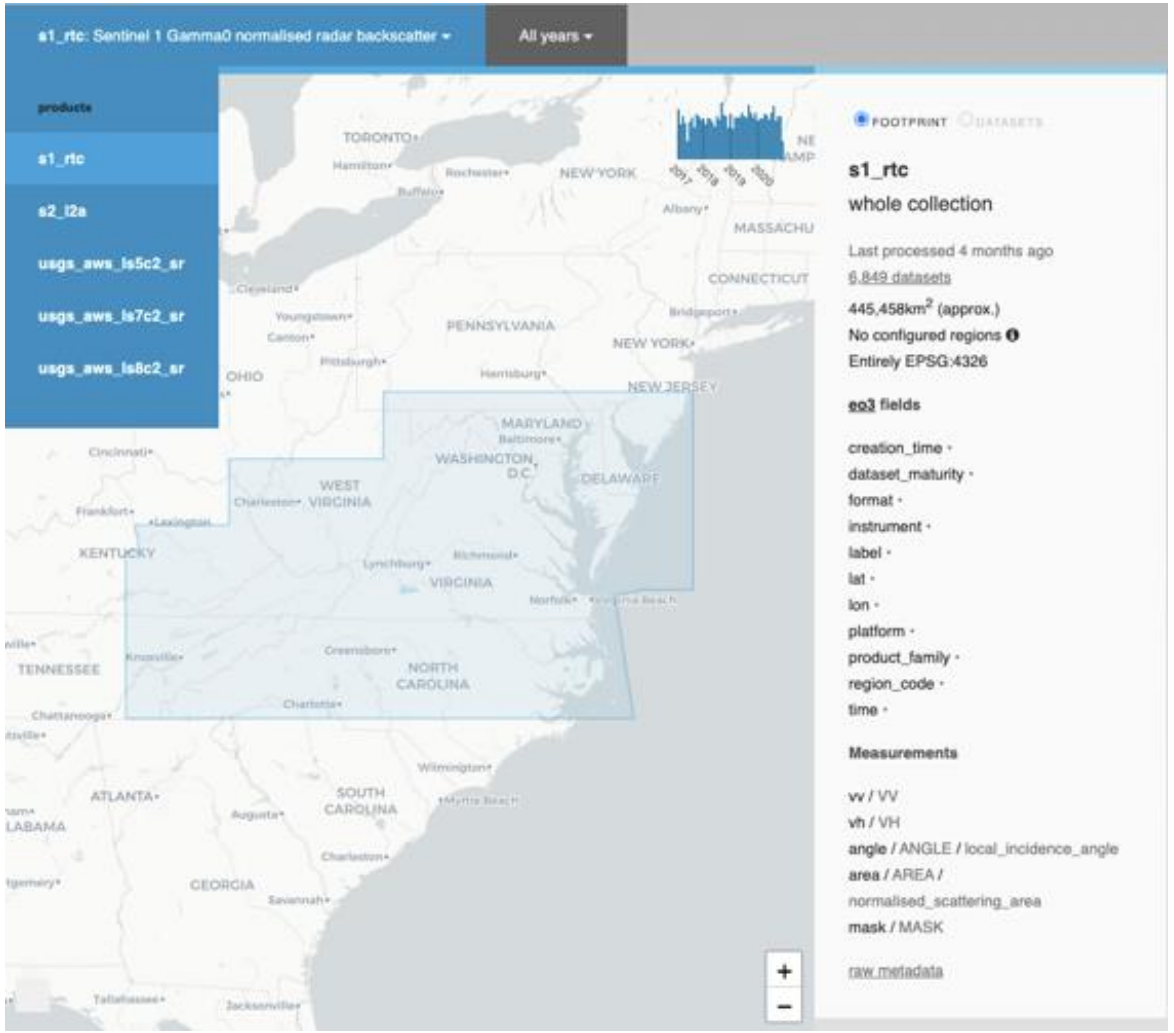
- CAL is a platform to serve the CEOS community and facilitate collaborative Earth Observation analysis
- Objectives:
  - Provide an open Jupyter notebook environment
  - Simplify loading of Earth observation data
  - Minimize analysis setup time and overhead

CAL Activities

- **CAL** currently supports the **COAST** (Chesapeake Bay study) project for CEOS and working with the CEOS **Ecosystem Extent** Pilot Project

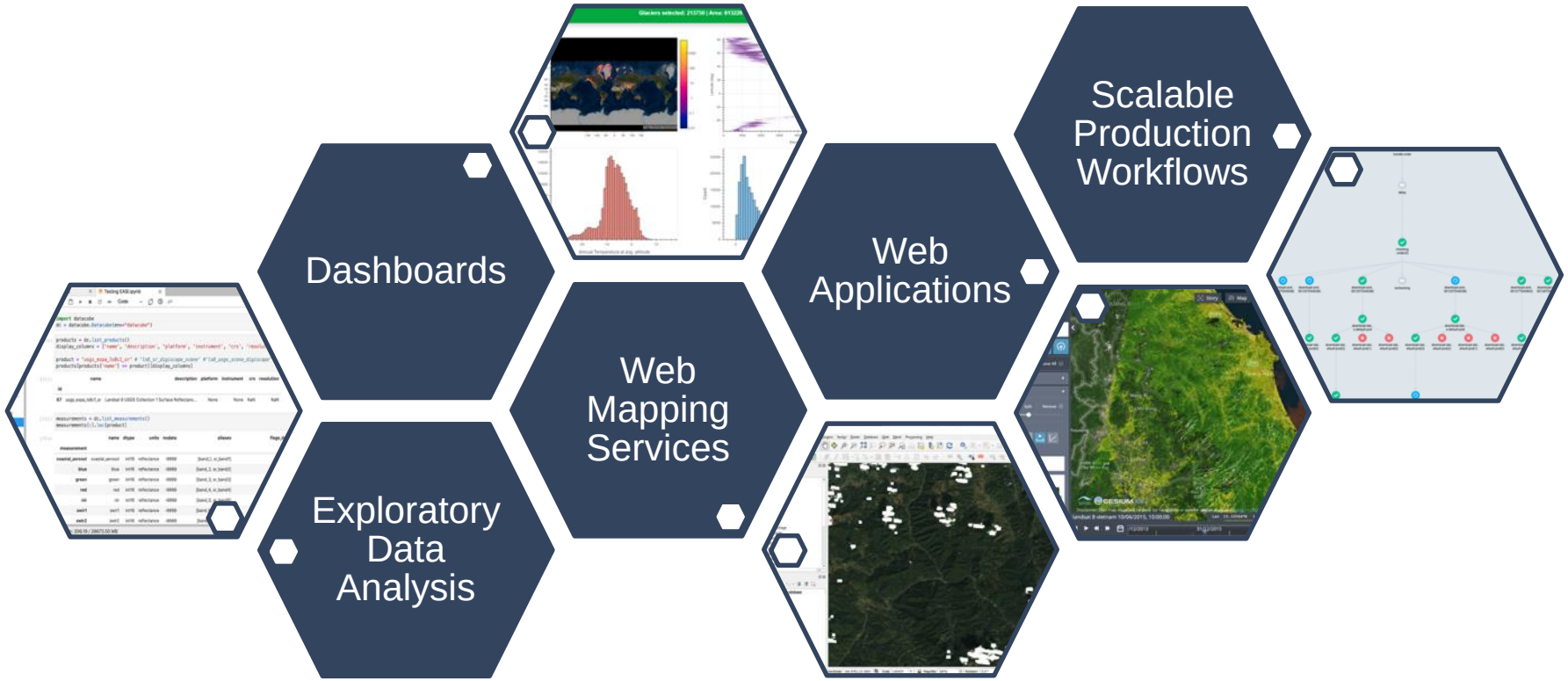
- Potential future users include: **WGCV** (DEMIX Cal-Val campaigns), **DE-Americas** (Caribbean Pilot project), and CEOS **Ecosystem Extent** Pilot Project

- **Datasets** currently available: Landsat, Sentinel-2, MODIS, Sentinel-3, Sentinel-1 (CARD4L with RTC), Copernicus DEM, and NASA DEM



Architecture and Services

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



OPEN DATA CUBE

jupyter

terria

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

aws + NVIDIA

jupyter + R

scikit learn

TensorFlow

dask

Keras

PyTorch

scikit-image
image processing in python

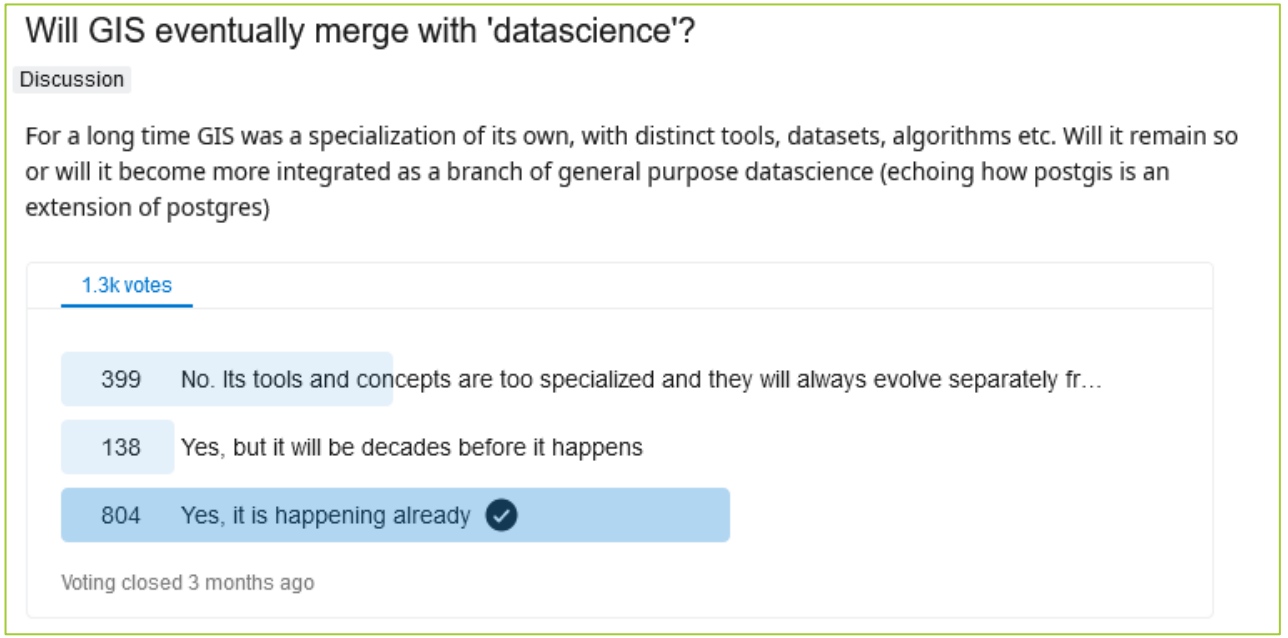
dmlc XGBoost

TorchGeo

Eli5

Data Science and GIS are Merging

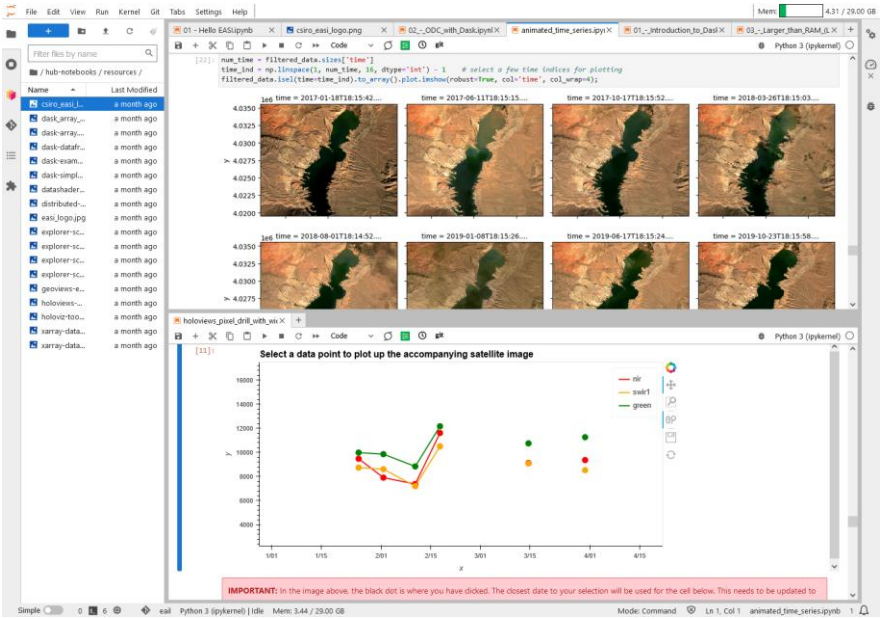
- Data science tools are being used to tackle big EO data
- Notebooks are a ubiquitous data science tool
- Jupyter notebooks are web based interactive development environment where textual explanations, graphical outputs and code are presented together



- Enables users to conduct analyses with common Python data science tools

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

Interoperability

- 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

Challenges

- 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

Support

- 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

CSIRO Chile

Committee on Earth Observation Satellites

DATA OBSERVATORY