IVELINA MOMCHEVA, STSCI

MULTI-MESSENGER ASTRONOMY IN THE CLOUD

IVELINA MOMCHEVA, STSCI

(MULTI-MESSENGER) ASTRONOMY IN THE CLOUD

LANDSCAPE: LOTS OF DATA

- LSST, Euclid, WFIRST producing large surveys
- LSST time domain: 10^7 events per night
- Dramatic improvement of MMA capabilities in the 2020s and 2030s: multiple events per day
- “Cornucopia” both in data scale and rate of acquisition
- Challenge: analysis over different messengers, many different data sources

STATUS QUO

ESO

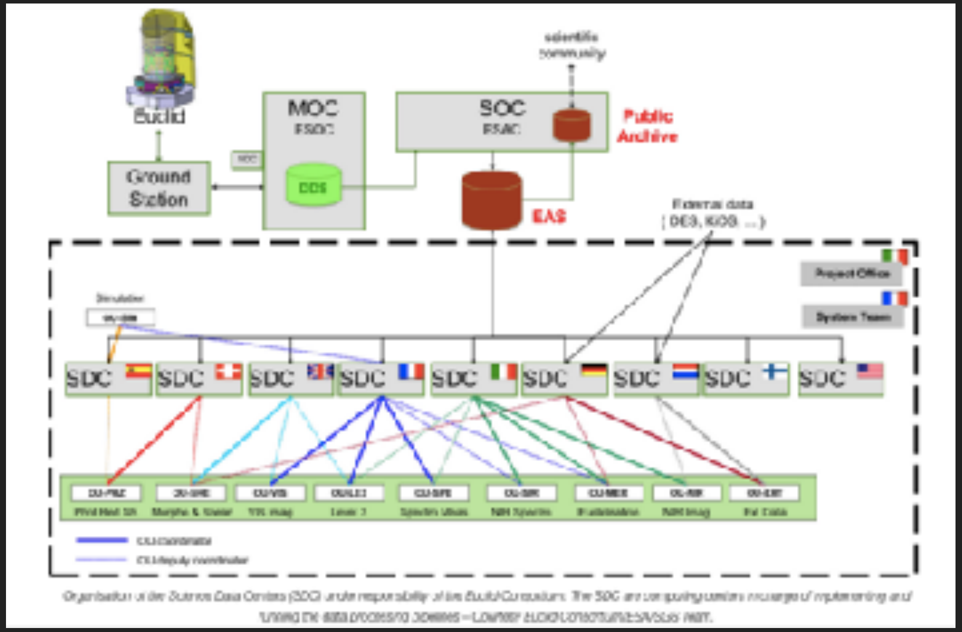
esa

NOAO DATALAB

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

LSST Large Synoptic Survey Telescope

ipac



++ many others.

> Cyberinfrastructure will need to be designed, built, and operated in order to collect, process, analyze, synthesize, and enable interpretation of time-critical and archival data from observing facilities across the multi-messenger spectra. Robust tools for data sharing, collaboration, joint analysis, as well as training and education will also need to be developed.

Chang et al., 2019
“Cyberinfrastructure Requirements to Enhance Multi-messenger Astrophysics ”
https://scimma.org

CHALLENGES WITH THE STATUS QUO

- Physical infrastructure: Infrastructure purchases made every ~5 years. Major modifications only possible during procurement/upgrade processes.
- Lack of standards: Each archive implements similar (but not quite the same!) solutions to common needs.
- Distributed architecture: data is geographically spread out, simultaneous access across datasets or of the full dataset impossible
- Hard to evolve data access model: 'Rigidity' of infrastructure means that experimenting with novel technology/new data management is very difficult.

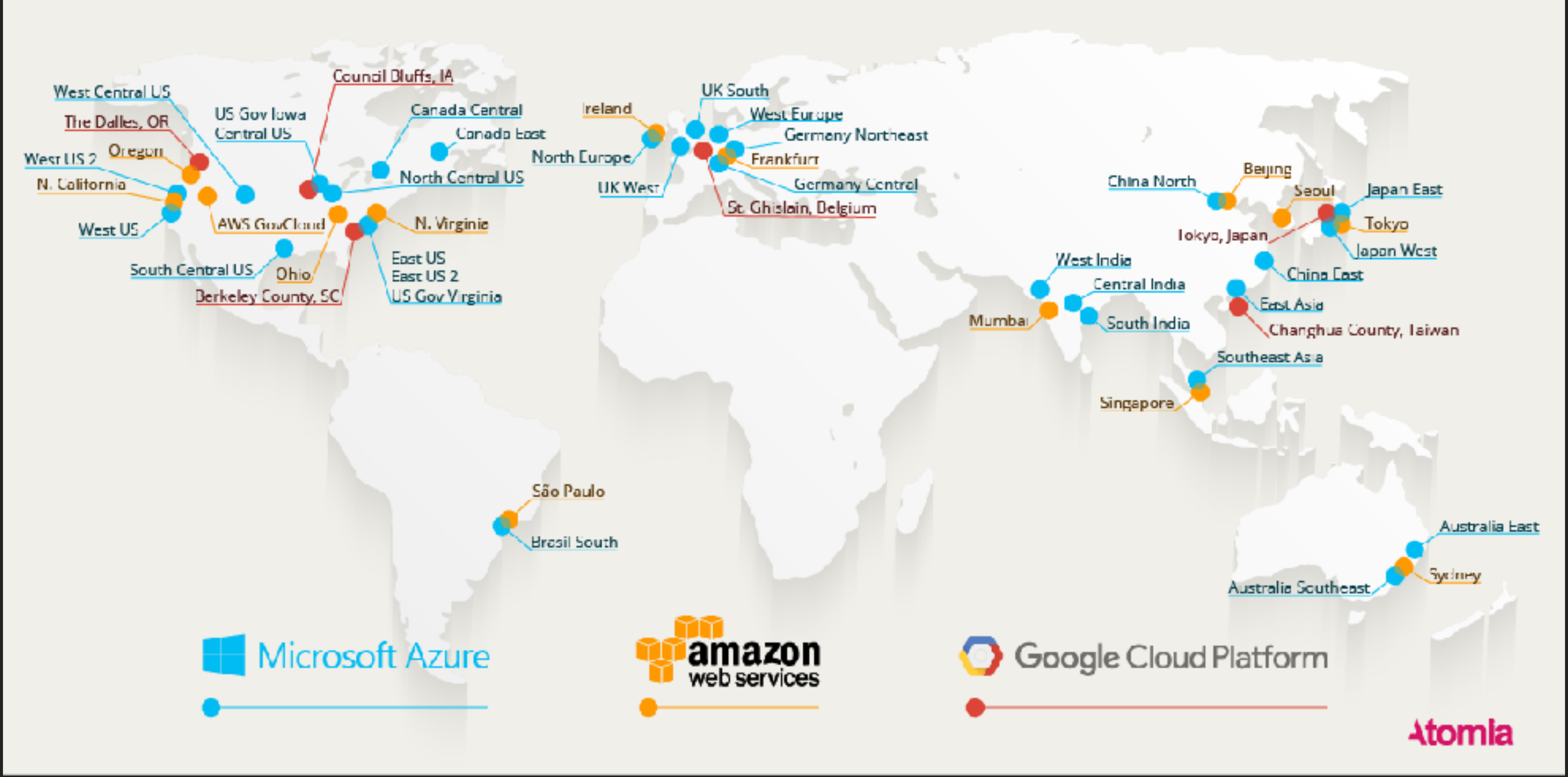
CLOUD COMPUTING: WHAT IS IT?

- Cloud = data center which provides on-demand compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing to offer faster innovation, flexible resources, and economies of scale

- Major benefits: on-demand self-service, fast network access, geographic distribution, resource pooling, rapid elasticity, reliable service

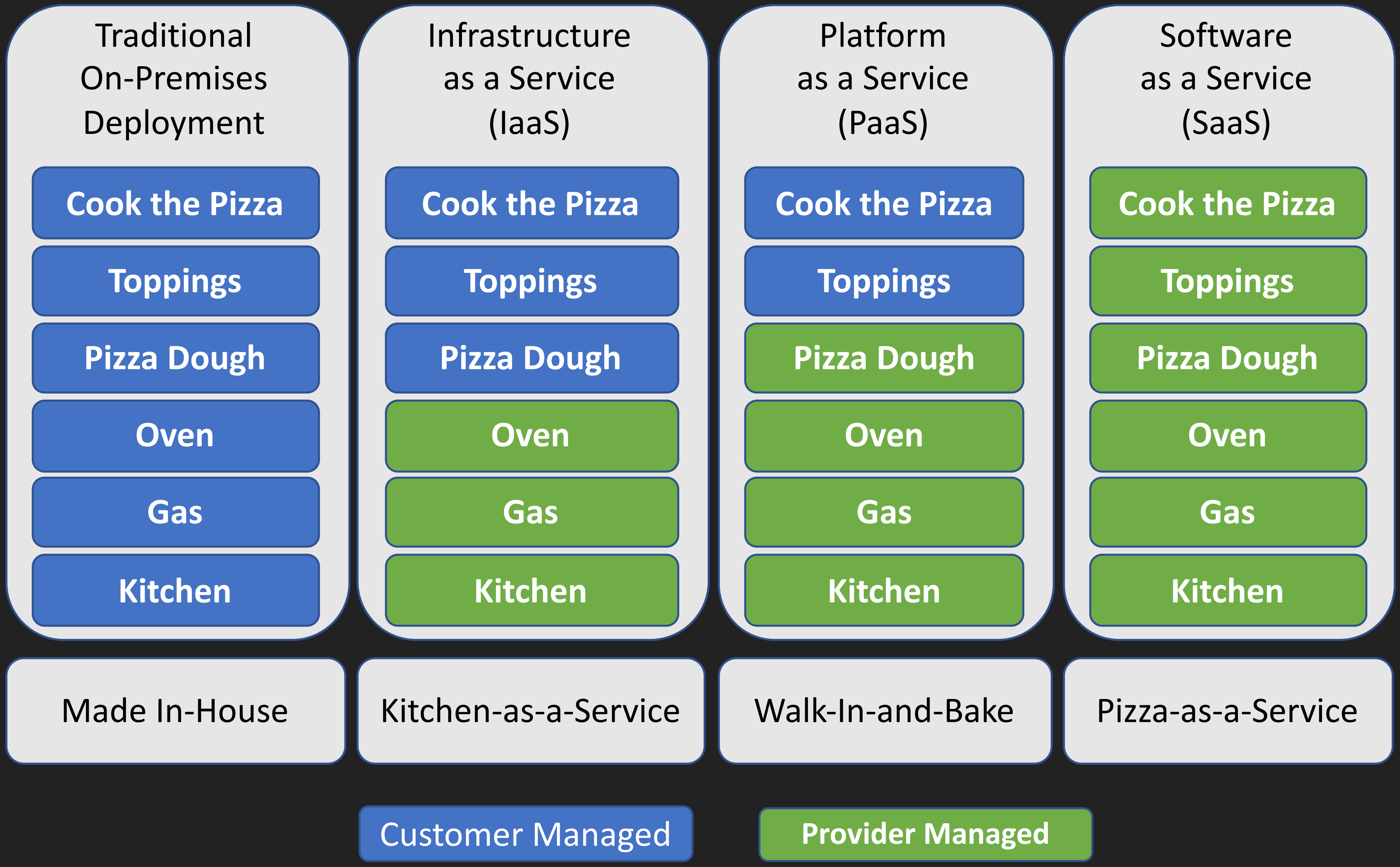
CLOUD COMPUTING: WHAT IS IT?

CLOUD COMPUTING: WHERE IS IT?



as of 2016

CLOUD COMPUTING: MODELS OF RESPONSIBILITY



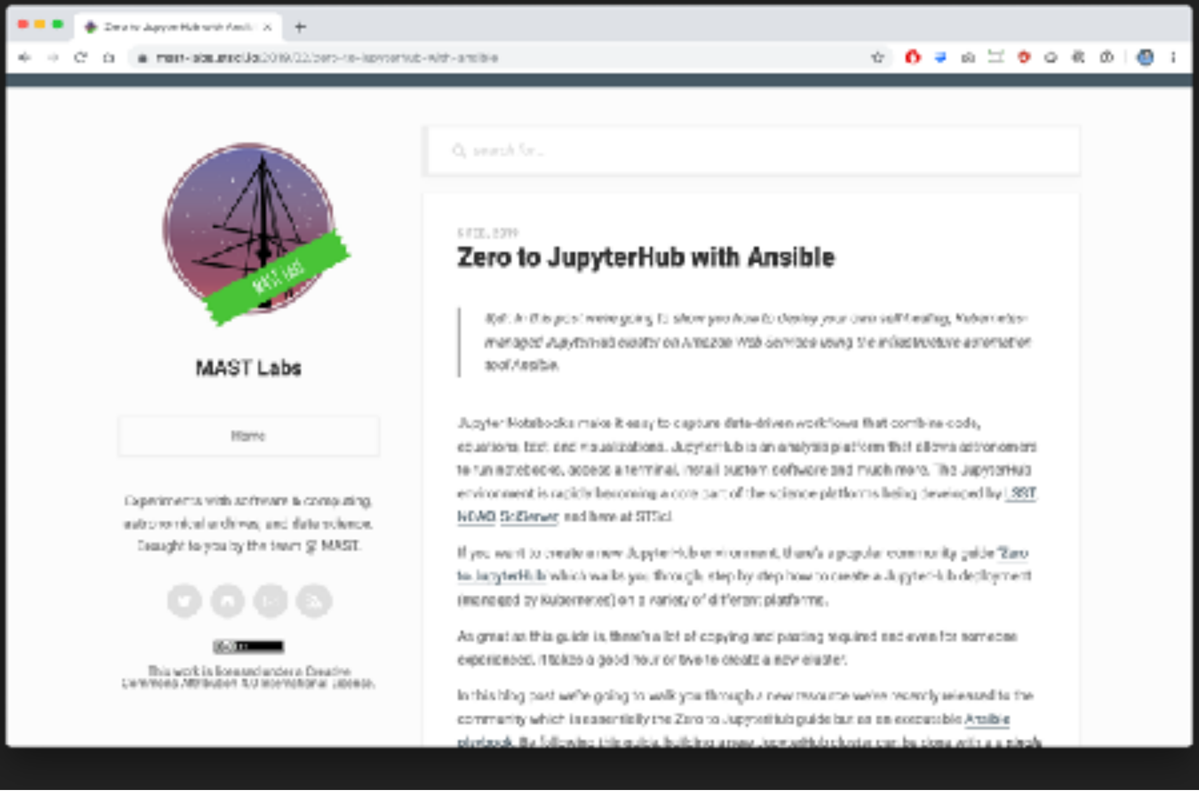
Traditional On-Premises Deployment	Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
🟦 Cook the Pizza	🟦 Cook the Pizza	🟦 Cook the Pizza	🟩 Cook the Pizza
🟦 Toppings	🟦 Toppings	🟦 Toppings	🟩 Toppings
🟦 Pizza Dough	🟦 Pizza Dough	🟩 Pizza Dough	🟩 Pizza Dough
🟦 Oven	🟩 Oven	🟩 Oven	🟩 Oven
🟦 Gas	🟩 Gas	🟩 Gas	🟩 Gas
🟦 Kitchen	🟩 Kitchen	🟩 Kitchen	🟩 Kitchen
Made In-House	Kitchen-as-a-Service	Walk-In-and-Bake	Pizza-as-a-Service

🟦 Customer Managed

🟩 Provider Managed

CLOUD COMPUTING: INFRASTRUCTURE AS CODE (IaC)

- IaC allows users to describe arbitrarily complex IT infrastructure in the form of configuration files and machine-readable scripts that can be used to provision infrastructure on-demand in the cloud on a 'pay as you go' basis



mast-labs.stsci.io/2019/02/zero-to-jupyterhub-with-ansible

MAST Labs

Home

Experiments with software & computing, astronomical archives, and data science. Brought to you by the team @ MAST.

This work is licensed under a Creative Commons Attribution 4.0 International License.

[illegible]

Zero to JupyterHub with Ansible

> *tl;dr: In this post we're going to show you how to deploy your own self-healing, Kubernetes-managed JupyterHub cluster on Amazon Web Services using the infrastructure automation tool Ansible.*

Jupyter Notebooks make it easy to capture data-driven workflows that combine code, equations, text, and visualizations. JupyterHub is an analysis platform that allows astronomers to run notebooks, access a terminal, install custom software and much more. The JupyterHub environment is rapidly becoming a core part of the science platforms being developed by LSST, NOAO SciServer and here at STScI.

If you want to create a new JupyterHub environment, there's a popular community guide 'Zero to JupyterHub' which walks you through step by step how to create a JupyterHub deployment (managed by Kubernetes) on a variety of different platforms.

As great as this guide is, there's a lot of copying and pasting required and even for someone experienced, it takes a good hour or two to create a new cluster.

In this blog post we're going to walk you through a new resource we've recently released to the community which is essentially the Zero to JupyterHub guide but as an executable Ansible playbook. By following this guide building a new JupyterHub cluster can be done with a single

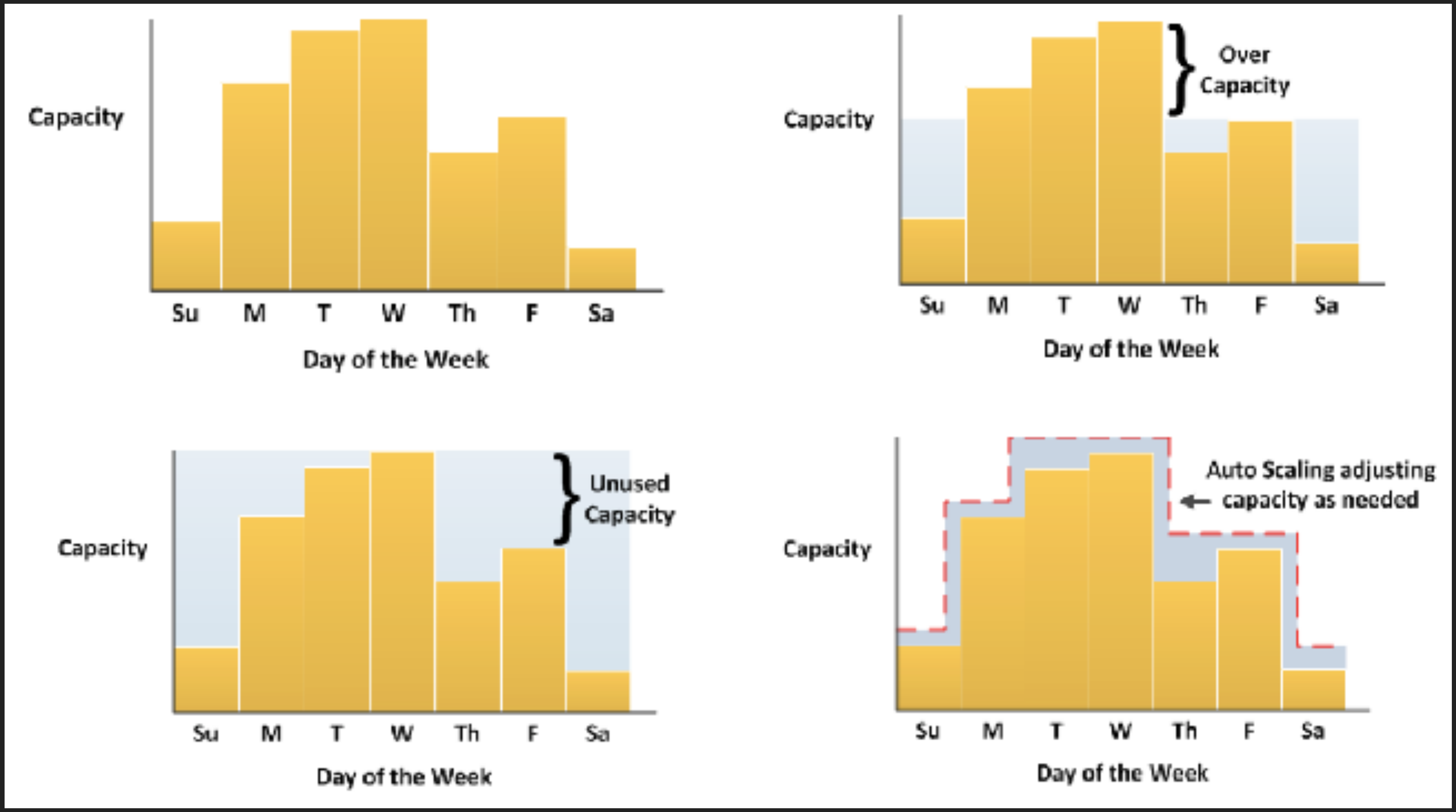
```
$ ansible-playbook -i hosts z2jh.yml
```

GENERIC CLOUD BENEFITS

- **On-demand**
- Elasticity
- Productivity
- Reliability
- Security
- Global scale
- Cost

GENERIC CLOUD BENEFITS

- On-demand
- **Elasticity**
- Productivity
- Reliability
- Security
- Global scale
- Cost



GENERIC CLOUD BENEFITS

- On-demand
- Elasticity
- **Productivity**
- Reliability
- Security
- Global scale
- Cost

GENERIC CLOUD BENEFITS

- On-demand
- Elasticity
- Productivity
- **Reliability**
- Security
- Global scale
- Cost

GENERIC CLOUD BENEFITS

- On-demand
- Elasticity
- Productivity
- Reliability
- **Security**
- Global scale
- Cost

GENERIC CLOUD BENEFITS

- On-demand
- Elasticity
- Productivity
- Reliability
- Security
- **Global scale**
- Cost

GENERIC CLOUD BENEFITS

- On-demand
- Elasticity
- Productivity
- Reliability
- Security
- Global scale
- **Cost**

CAPITAL EXPENSE
+
ONGOING MAINTENANCE

VS

VARIABLE EXPENSE

BENEFITS TO ASTRONOMICAL DATA MANAGEMENT

- Reuse infrastructure: focus on core competencies, experiment with new technologies
- Scale: work server-side, high-throughput computing and pleasingly parallel work possible
- Innovation and accessibility: lower barrier to entry, increase innovation, democratize compute access

BENEFITS TO ASTRONOMICAL DATA MANAGEMENT

- Reuse infrastructure: focus on core competencies, experiment with new technologies
- Scale: work server-side, high-throughput computing and pleasingly parallel work possible
- Innovation and accessibility: lower barrier to entry, increase innovation, democratize compute access

BENEFITS TO ASTRONOMICAL DATA MANAGEMENT

- Reuse infrastructure: focus on core competencies, experiment with new technologies
- Scale: work server-side, high-throughput computing and pleasingly parallel work possible
- Innovation and accessibility: lower barrier to entry, increase innovation, democratize compute access

PERSONNEL

- Adoption of common technologies can increase inter-project mobility
- Decrease single points of failure/software start-up factor
- Increase the recruitment pool

EDUCATION

- Developing cloud-relevant skills as part of astro training at all levels
- Providing career paths for people with the necessary skill set

Norman et al., 2019, “The Growing Importance of a Tech-Savvy Astronomy and Astrophysics Workforce ”

SOFTWARE DEVELOPMENT

- We can all share and contribute to fundamental software libraries
- The IaC paradigm allows for reuse of infrastructure components and cumulative development
- If you did something cool, push it upstream

POTENTIAL CHALLENGES

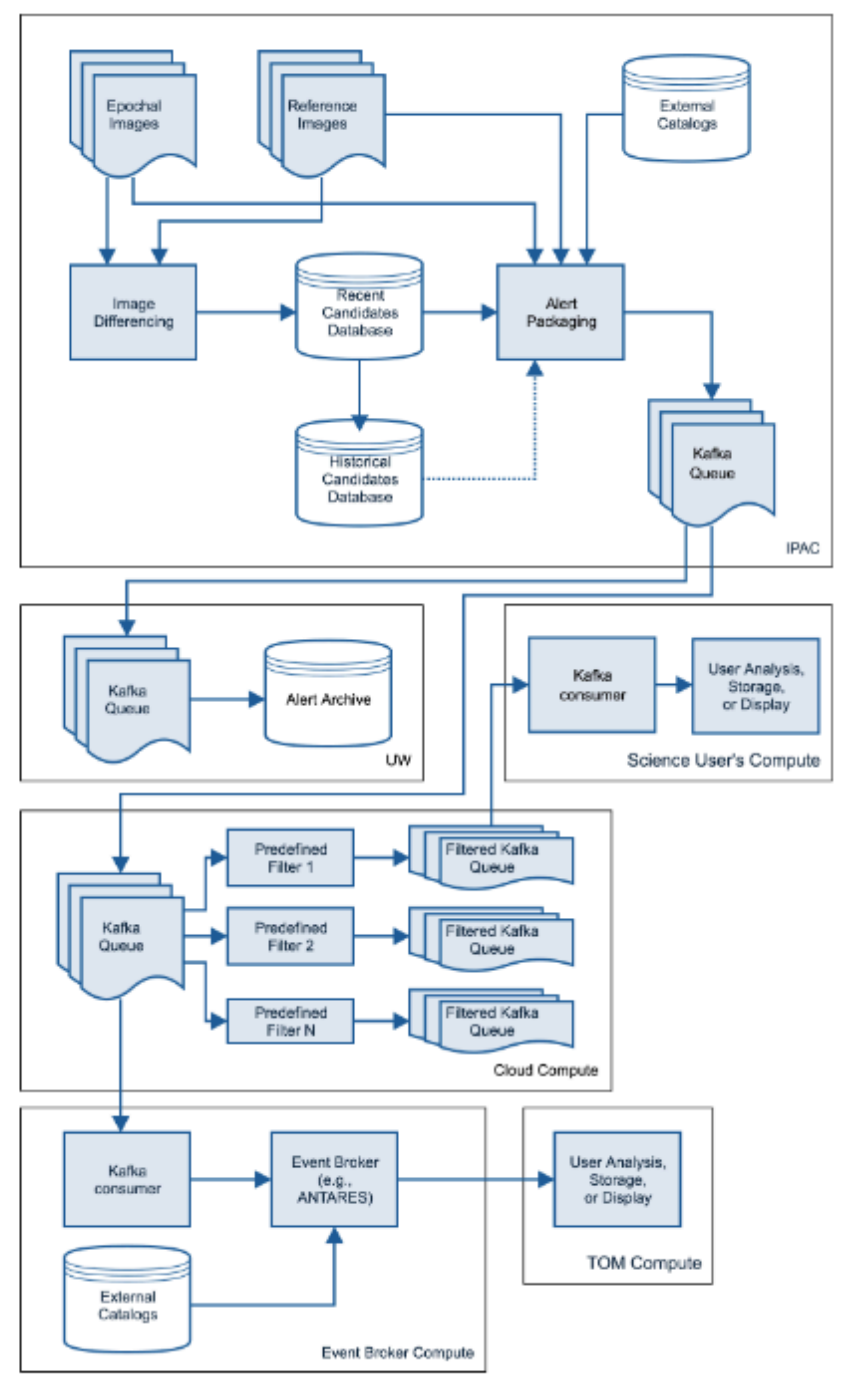
- Astronomical workflows need to be refactored for the cloud
- Storage costs and egress fees, running some services free for the community is non-negotiable
- Security: a *well managed* cloud can be more secure than “on-prem”
- Vendor lock-in and commercial partnerships

HOW CAN WE BENEFIT FROM CLOUD COMPUTING?

Time Domain and MMA Communities

SCIOPS-2019

- TD and MMA workflows very well suited to **event-driven architecture**: software architecture paradigm based on the production, detection, consumption to and reaction to **events**.
- Components are loosely coupled by publish/subscribe (**pub/sub**) messaging. Many different applications can be listening to the same stream.
- Many open source technologies exist for building event-driven systems: e.g., Apache Kafka, Apache Spark. Available as PaaS from AWS/Google.
- Event serialization
- Decorating streams
- Coupling with other cloud services: serverless compute (FaaS/Lambda/Cloud Functions), machine learning

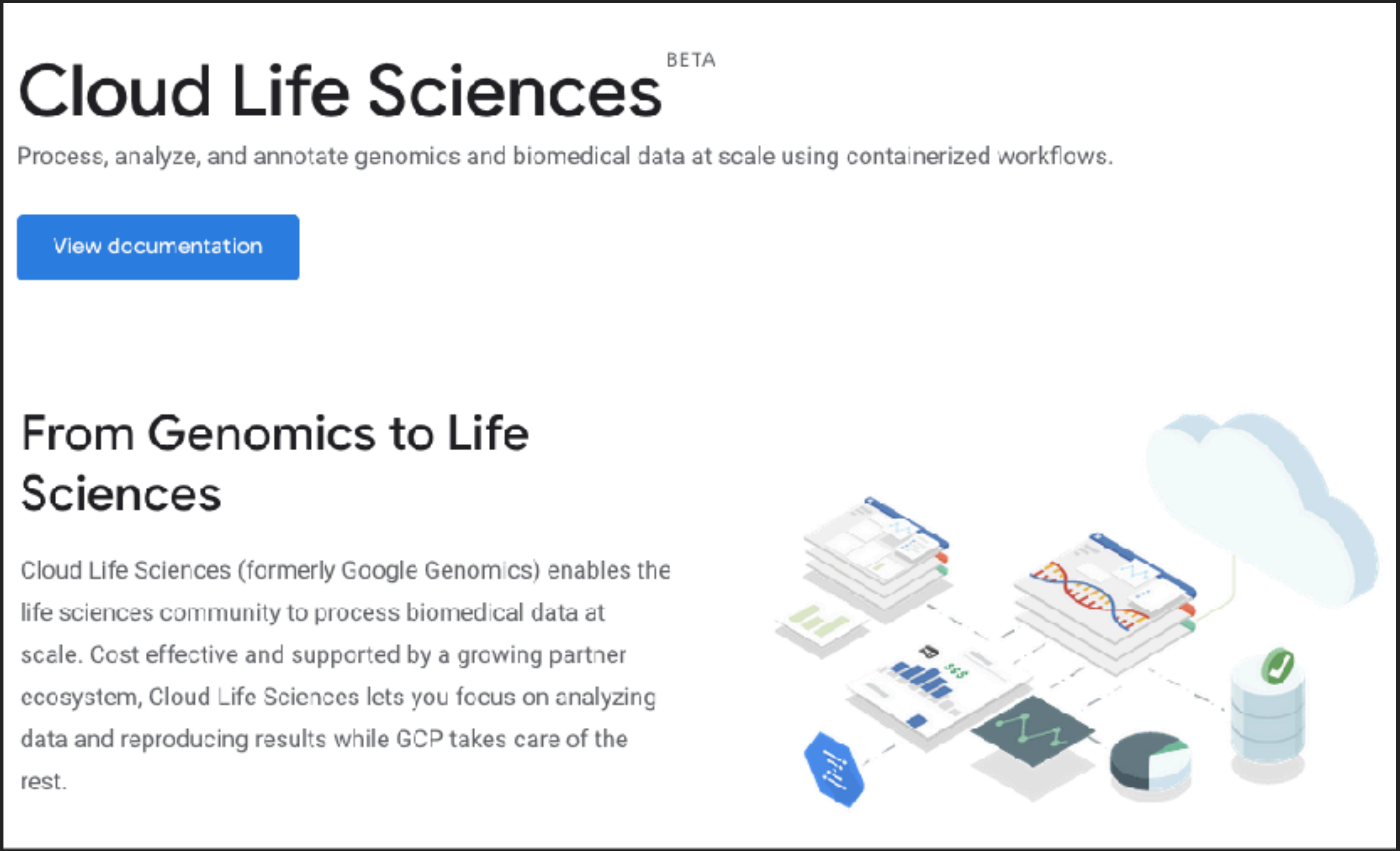


Patterson et al., 2019, ZTF prototype

> This ability to use highly-reliable data management infrastructure on demand means the astrophysics community can focus on its core competencies such as developing novel algorithms for extracting science from mission data rather than building and maintaining custom data management infrastructure which is often [..] more expensive than what is made available by [public cloud providers].

Smith et al., 2019
"Astronomy should be in the clouds"

CLOUD LIFE SCIENCES



Cloud Life Sciences BETA

Process, analyze, and annotate genomics and biomedical data at scale using containerized workflows.

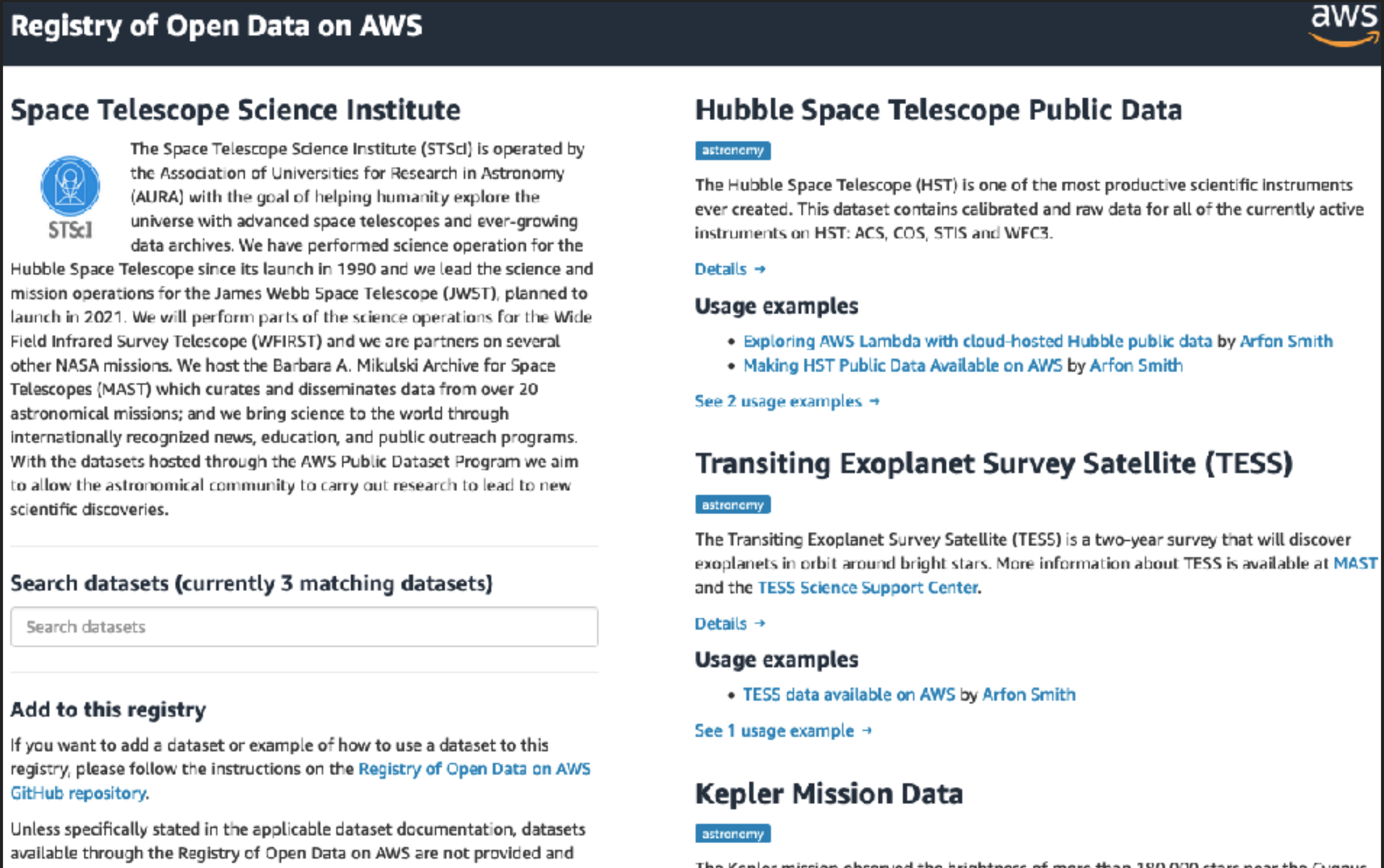
From Genomics to Life Sciences

Cloud Life Sciences (formerly Google Genomics) enables the life sciences community to process biomedical data at scale. Cost effective and supported by a growing partner ecosystem, Cloud Life Sciences lets you focus on analyzing data and reproducing results while GCP takes care of the rest.

https://cloud.google.com/life-sciences/

HUBBLE EXAMPLE

https://registry.opendata.aws/collab/stsci/



Registry of Open Data on AWS

Space Telescope Science Institute

The Space Telescope Science Institute (STScI) is operated by the Association of Universities for Research in Astronomy (AURA) with the goal of helping humanity explore the universe with advanced space telescopes and ever-growing data archives. We have performed science operation for the Hubble Space Telescope since its launch in 1990 and we lead the science and mission operations for the James Webb Space Telescope (JWST), planned to launch in 2021. We will perform parts of the science operations for the Wide Field Infrared Survey Telescope (WFIRST) and we are partners on several other NASA missions. We host the Barbara A. Mikulski Archive for Space Telescopes (MAST) which curates and disseminates data from over 20 astronomical missions; and we bring science to the world through internationally recognized news, education, and public outreach programs. With the datasets hosted through the AWS Public Dataset Program we aim to allow the astronomical community to carry out research to lead to new scientific discoveries.

Search datasets (currently 3 matching datasets)

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the Registry of Open Data on AWS GitHub repository.

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and

Hubble Space Telescope Public Data

astronomy

The Hubble Space Telescope (HST) is one of the most productive scientific instruments ever created. This dataset contains calibrated and raw data for all of the currently active instruments on HST: ACS, COS, STIS and WFC3.

Details →

Usage examples

- Exploring AWS Lambda with cloud-hosted Hubble public data by Arfon Smith
- Making HST Public Data Available on AWS by Arfon Smith

See 2 usage examples →

Transiting Exoplanet Survey Satellite (TESS)

astronomy

The Transiting Exoplanet Survey Satellite (TESS) is a two-year survey that will discover exoplanets in orbit around bright stars. More information about TESS is available at MAST and the TESS Science Support Center.

Details →

Usage examples

- TESS data available on AWS by Arfon Smith

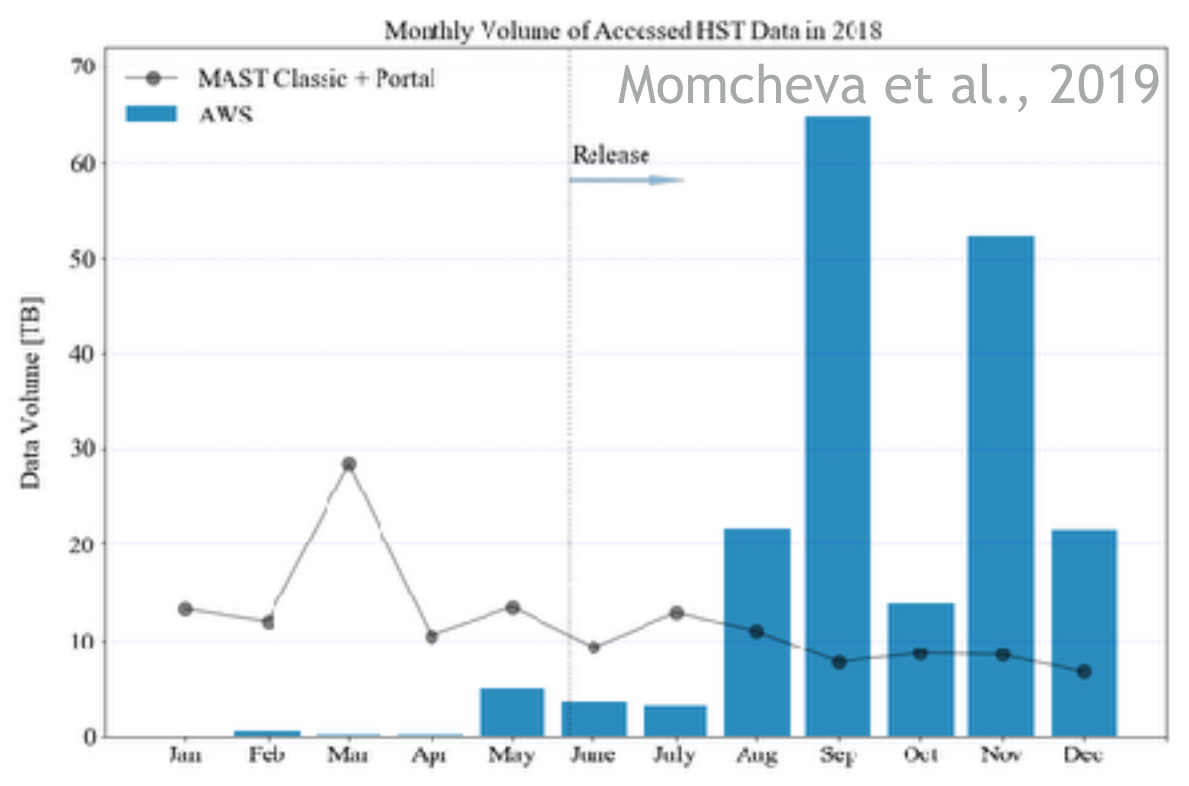
See 1 usage example →

Kepler Mission Data

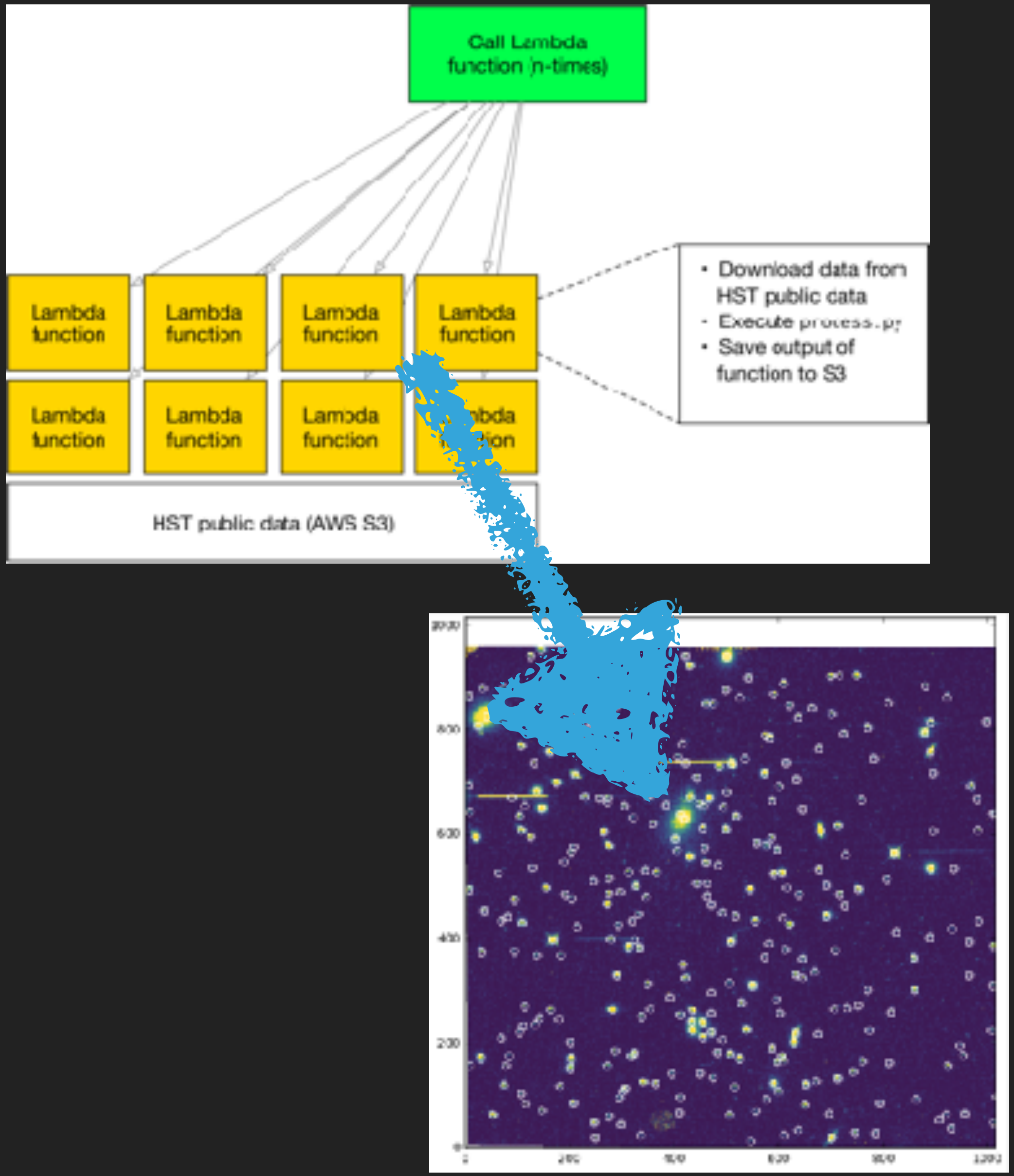
astronomy

The Kepler mission observed the brightness of more than 180,000 stars near the Cygnus

HUBBLE EXAMPLE



Momcheva et al., 2019



Call Lambda function (n-times)

Lambda function

Lambda function

Lambda function

Lambda function

Lambda function

Lambda function

Lambda function

Lambda function

- Download data from HST public data
- Execute process.py
- Save output of function to S3

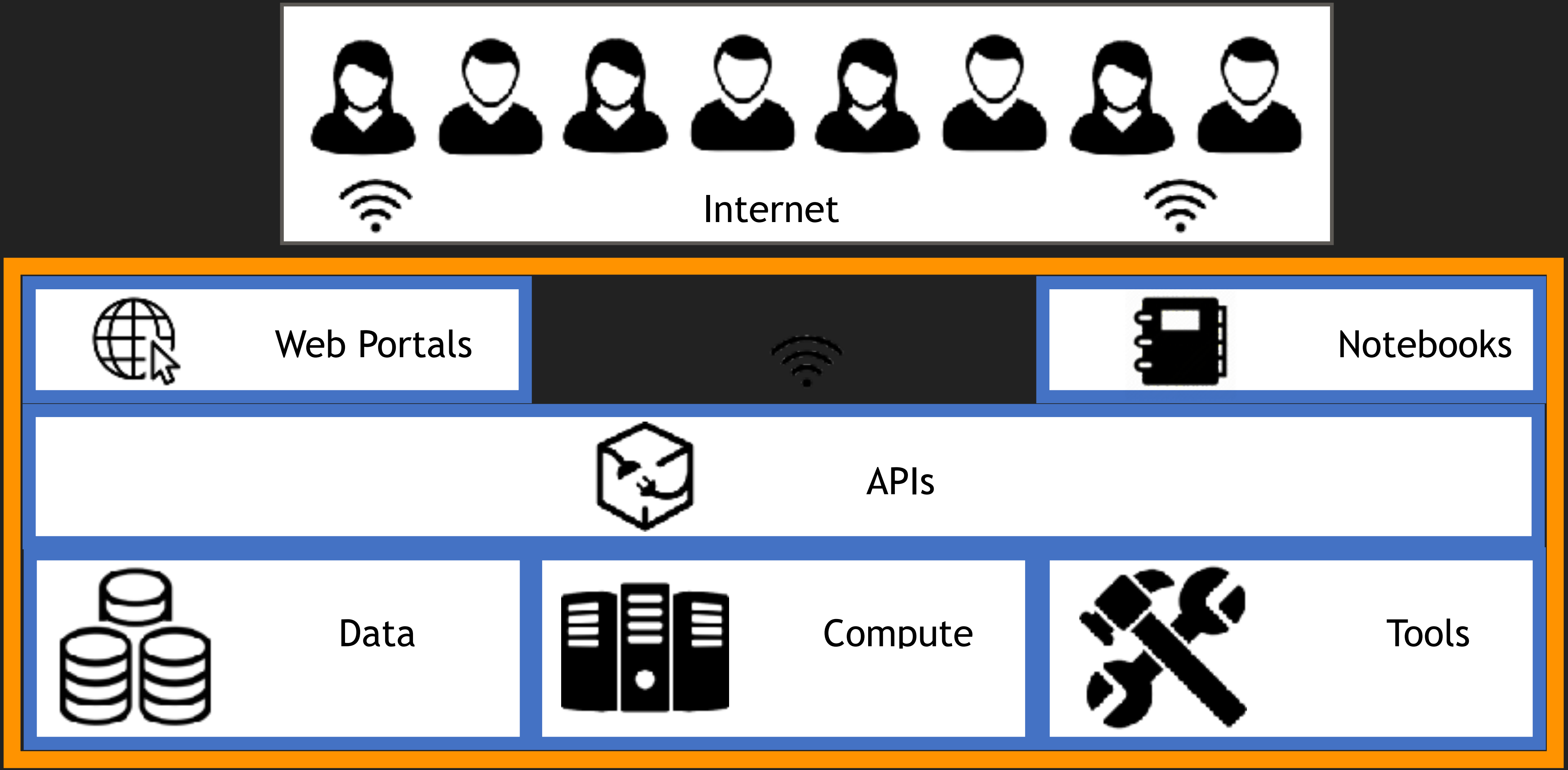
HST public data (AWS S3)

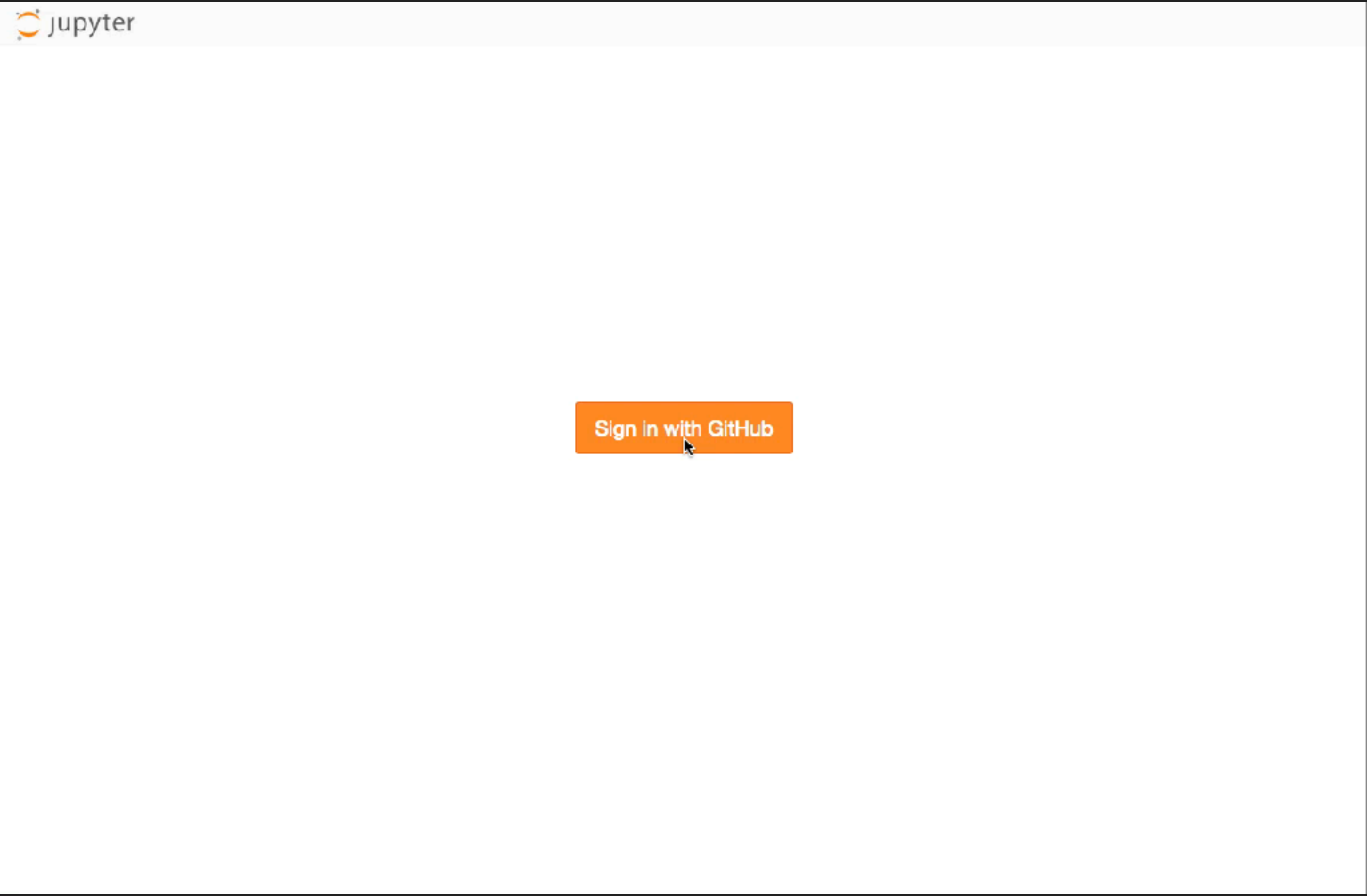
- Lambda all the HST images
- ML all the new HST images
- Run the HST pipeline in the cloud
- Correct astrometry to GAIA
- Reduce all HST grism data

https://mast-labs.stsci.io/

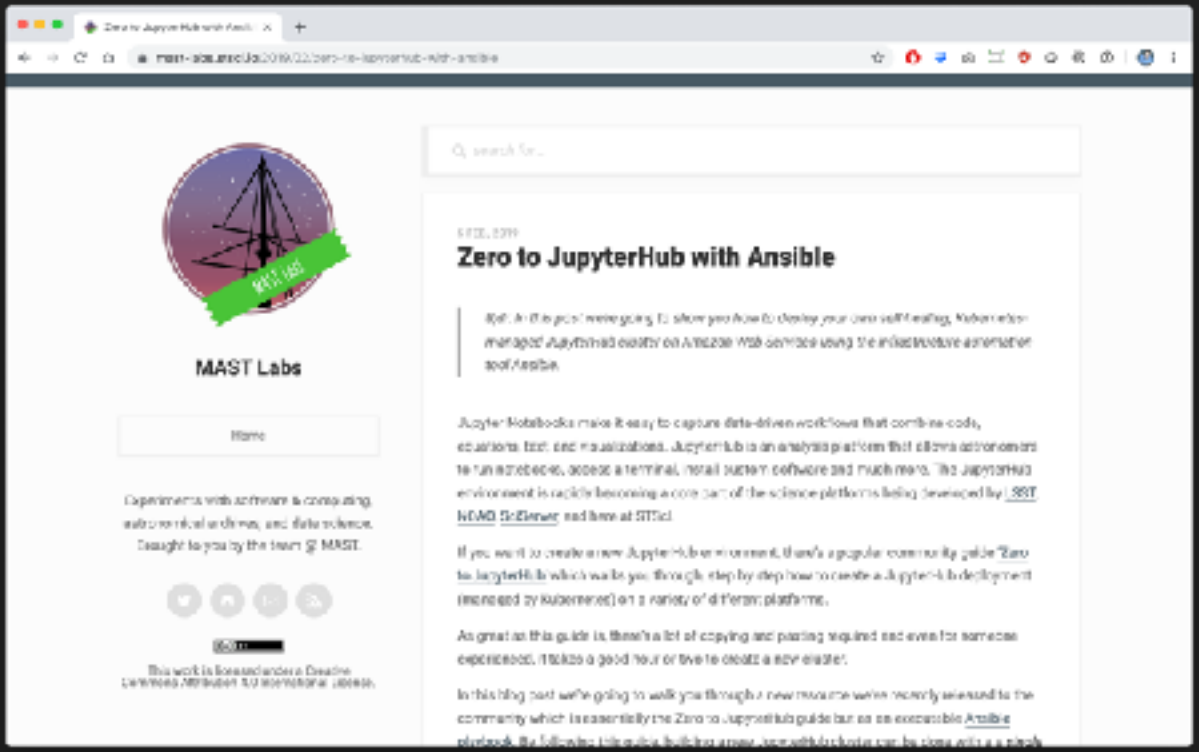
SERVER SIDE ANALYTICS/SCIENCE PLATFORMS

A **Science Platform** is an environment which combines data storage, computational capabilities, software tools and interfaces for users to interact with the underlying components.




jupyter
Sign in with GitHub

SERVER SIDE ANALYTICS/SCIENCE PLATFORMS



Zero to JupyterHub with Ansib

mast-labs.stsci.io/2019/02/zero-to-jupyterhub-with-ansible

search for...

MAST Labs

Home

Experiments with software & computing, astronomical archives, and data science. Brought to you by the team @ MAST.

[illegible] 2019

Zero to JupyterHub with Ansible

> *tl;dr: In this post we're going to show you how to deploy your own self-healing, Kubernetes-managed JupyterHub cluster on Amazon Web Services using the infrastructure automation tool Ansible.*

Jupyter Notebooks make it easy to capture data-driven workflows that combine code, equations, text, and visualizations. JupyterHub is an analysis platform that allows astronomers to run notebooks, access a terminal, install custom software and much more. The JupyterHub environment is rapidly becoming a core part of the science platforms being developed by LSST, NOAO SciServer and here at STScI.

If you want to create a new JupyterHub environment, there's a popular community guide 'Zero to JupyterHub' which walks you through step by step how to create a JupyterHub deployment (managed by Kubernetes) on a variety of different platforms.

As great as this guide is, there's a lot of copying and pasting required and even for someone experienced, it takes a good hour or two to create a new cluster.

In this blog post we're going to walk you through a new resource we've recently released to the community which is essentially the Zero to JupyterHub guide but as an executable Ansible playbook. By following this guide building a new JupyterHub cluster can be done with a single

docker

kubernetes

amazon webservices™

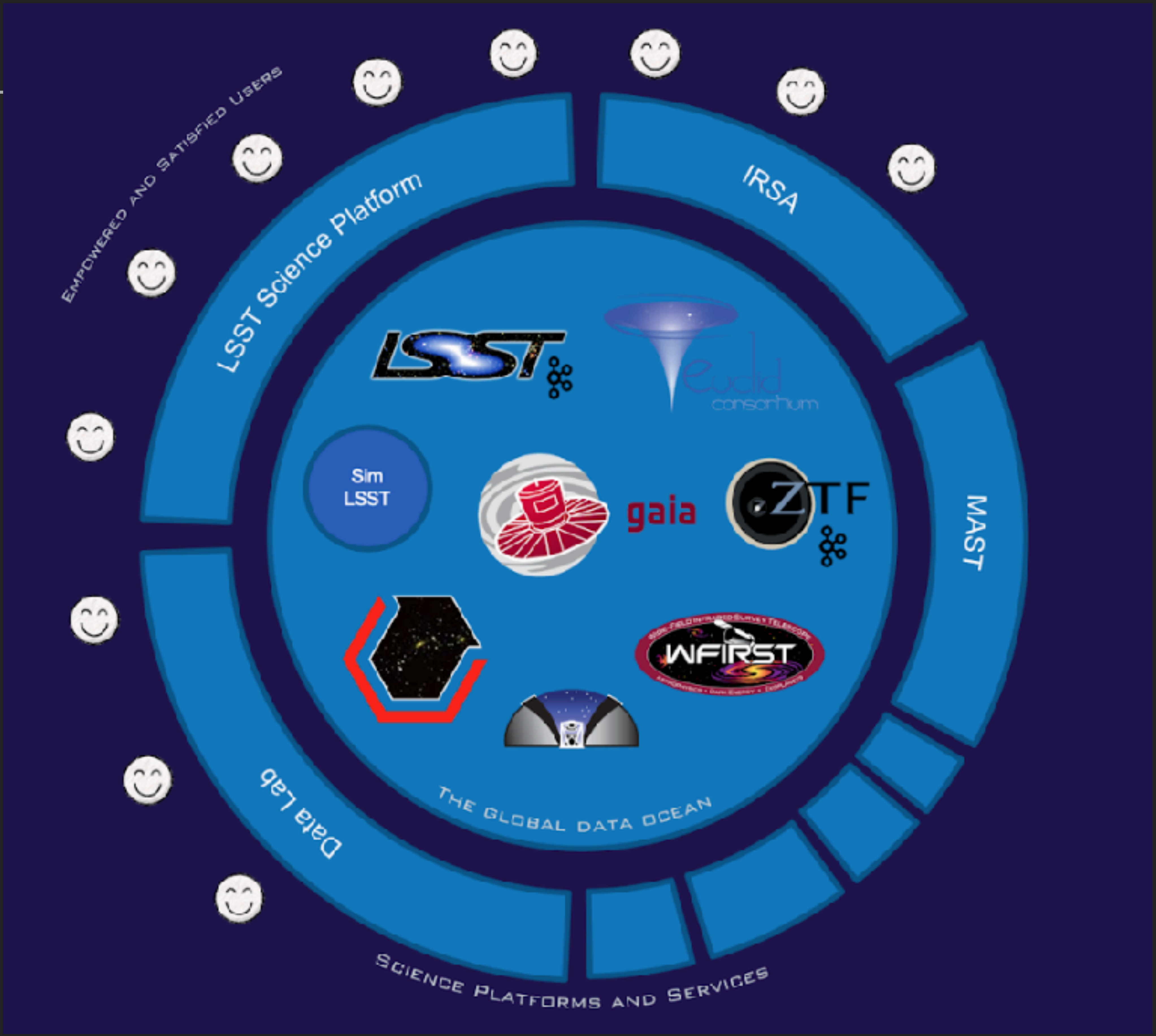
jupyterhub

```
$ ansible-playbook -i hosts z2jh.yml
```

https://mast-labs.stsci.io/2019/02/zero-to-jupyterhub-with-ansible

see Momcheva et al., 2019 ADASS proceeding

THE FUTURE?


EMPOWERED AND SATISFIED USERS
LSST Science Platform
IRSA
MAST
Data Lab
LSST
euclid consortium
Sim LSST
gaia
ZTF
WFIRST
THE GLOBAL DATA OCEAN
SCIENCE PLATFORMS AND SERVICES

Computing and storage are becoming commodities. Astronomical facilities should be moving away from custom infrastructure deployments.

Smith et al., 2019
“Astronomy should be in the clouds ”

CONCLUSIONS

- Cloud computing is a major trend in industry
- Cloud adoption can be completely opaque to astro users (the Netflix experience)
- A coordinated approach to cloud computing can usher in a new era in astronomical data management analysis
- New opportunities for data exploration will lead to new science

CLOUD VS ON-PREMISE PRICING

- On-premise expenses:
  - up-front hardware including redundancy
  - facilities and utilities (power, network)
  - IT staff to maintain and manage
  - ongoing refresh of hardware

- Cloud expenses:
  - pay-as-you go for using storage, compute, databases and services
  - multiple tiers depending on availability needs and work requirements
  - variable, negotiable and decreasing

- Summary: The cloud allows to trade capital expense (data centers, physical servers, etc.) for variable expense and only pay for IT as you consume it. Plus, the variable expense is much lower than what customers can do on their own because of the larger economies of scale.