Science Platforms

Ivelina Momcheva

Mission Scientist
Data Science Mission Office, STScI

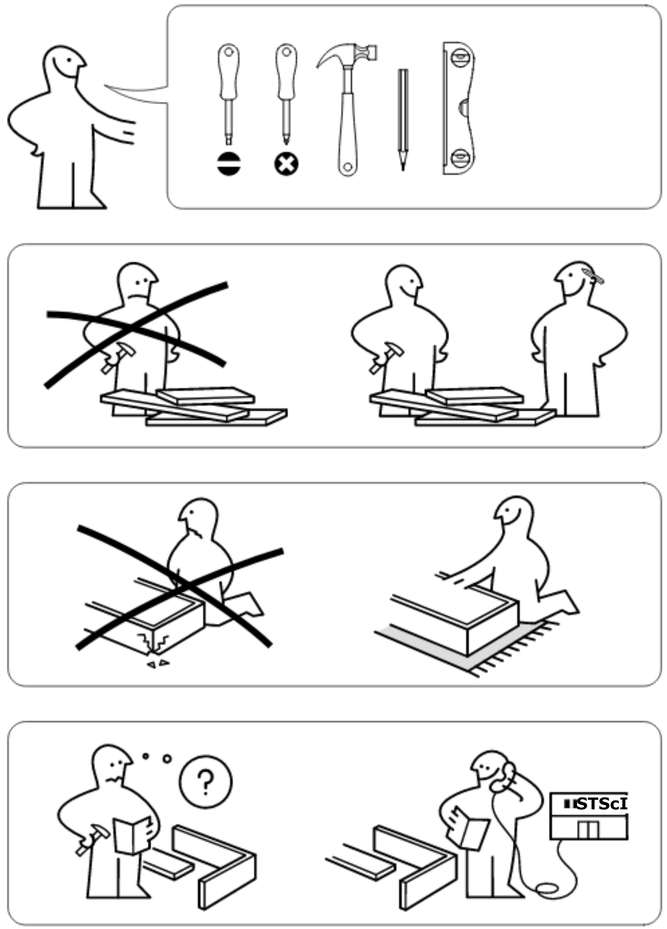
+ Arfon Smith, Josh Peek, Mike Fox
+ Jacob Matuskey, Christian Mesh, Erik Tollerud, Steve Crawford, DMD

Talk Overview

- How We Work with Data
- Science Platforms for Fun and Profit
- DIY Science Platforms
- Challenges and Future Directions

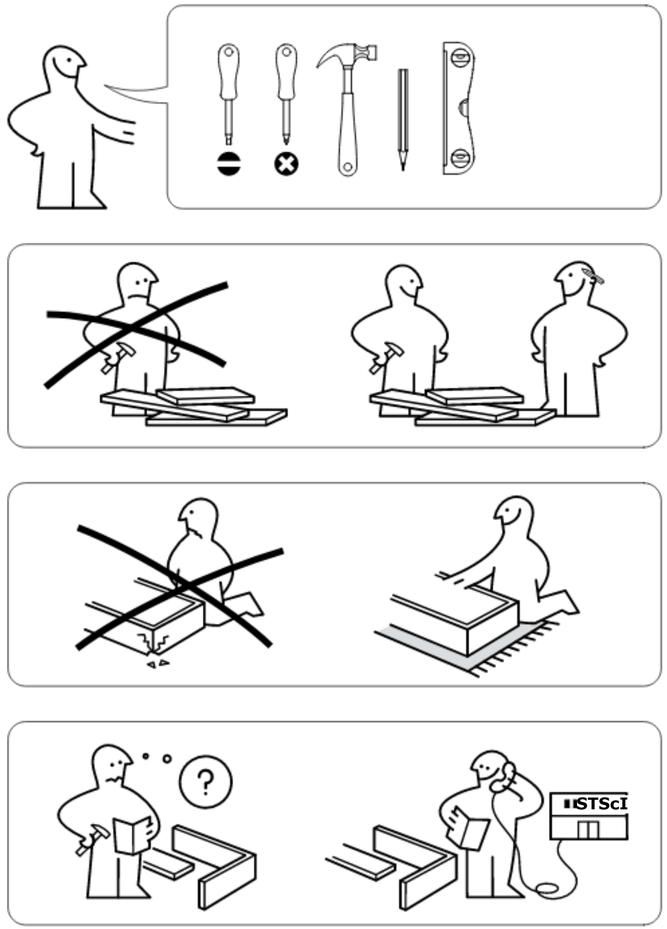
How Do We Process Data?

DÅTÅ PRØCESSING



How Do We Process Data?

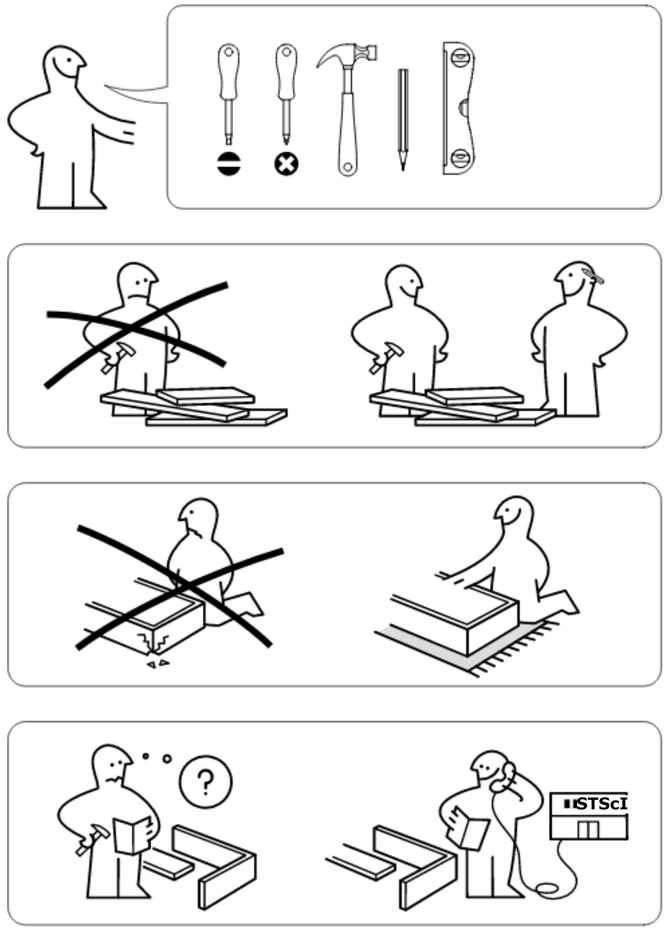
DÅTÅ PRØCESSING



1. INSTALL SØFTWÅRE

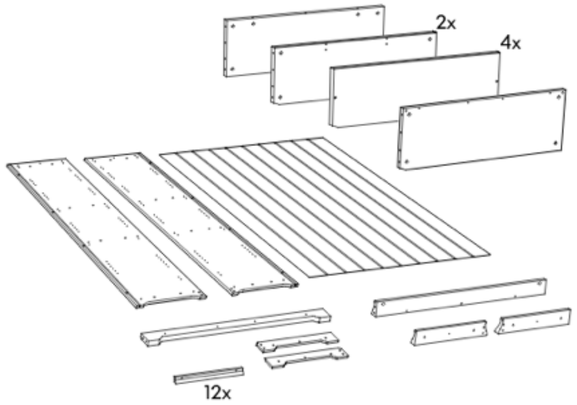
How Do We Process Data?

DÅTÅ PRØCESSING



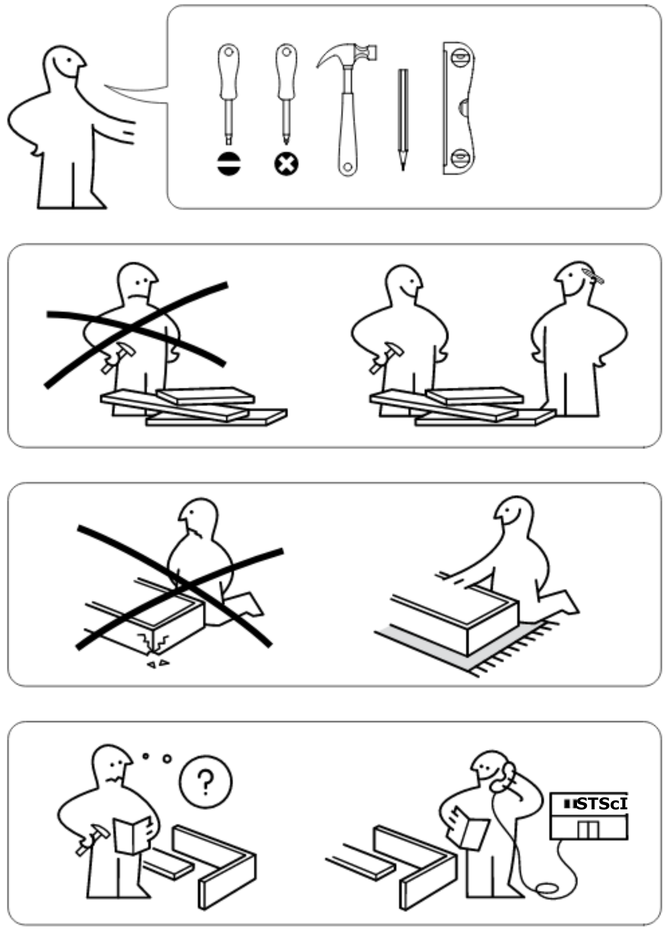
1. INSTALL SØFTWÅRE

2. GET DÅTÅ



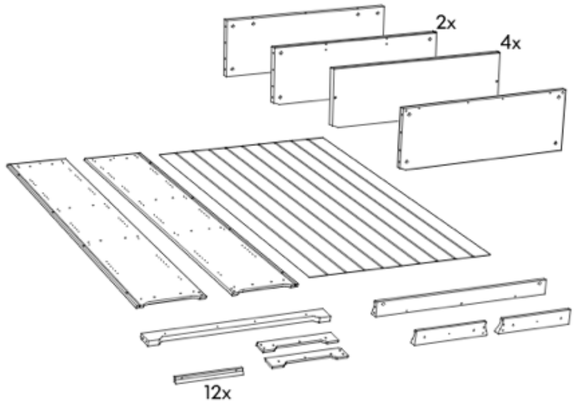
How Do We Process Data?

DÅTÅ PRØCESSING

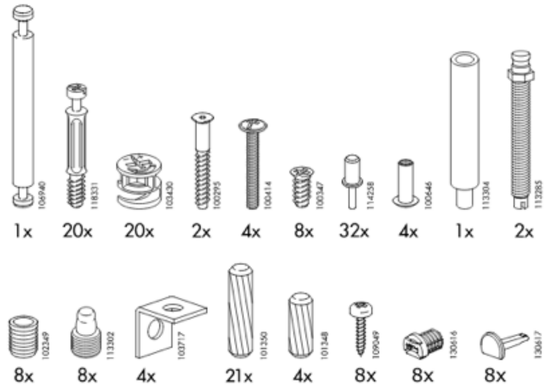


1. INSTALL SØFTWÅRE

2. GET DÅTÅ

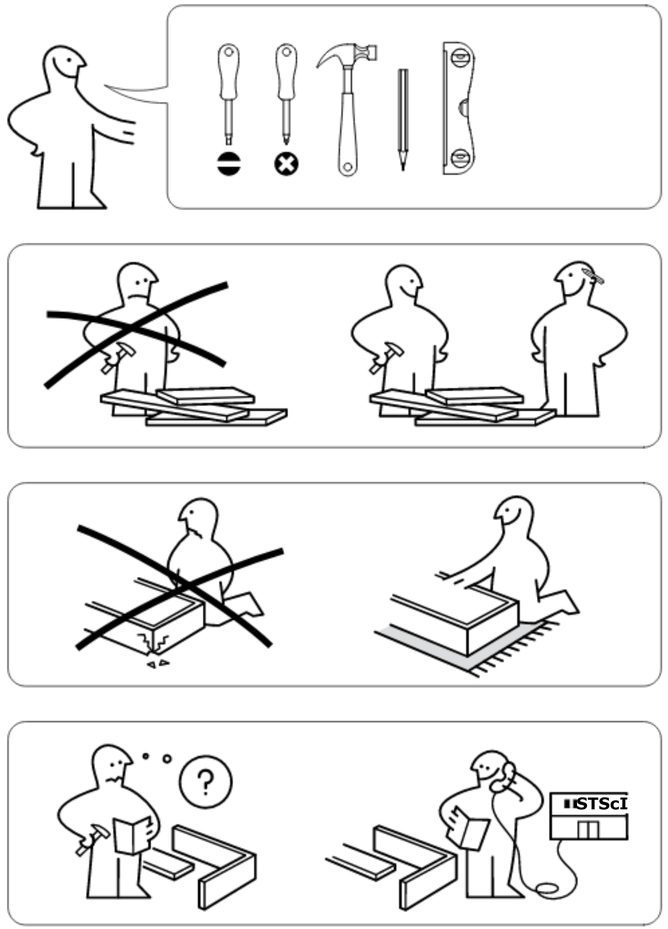


3. GET REFERENCE FILES



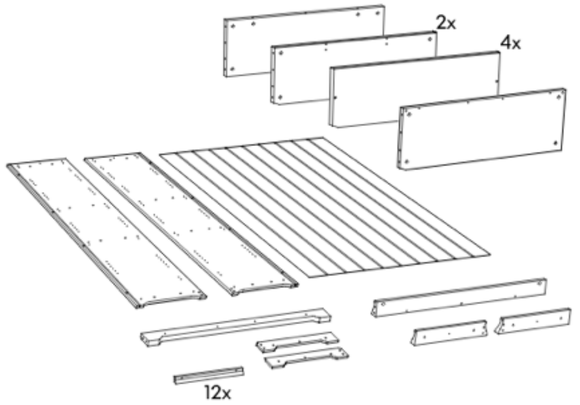
How Do We Process Data?

DÅTÅ PRØCESSING

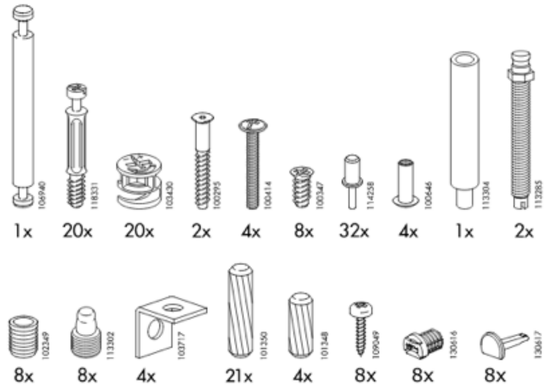


1. INSTALL SØFTWÅRE

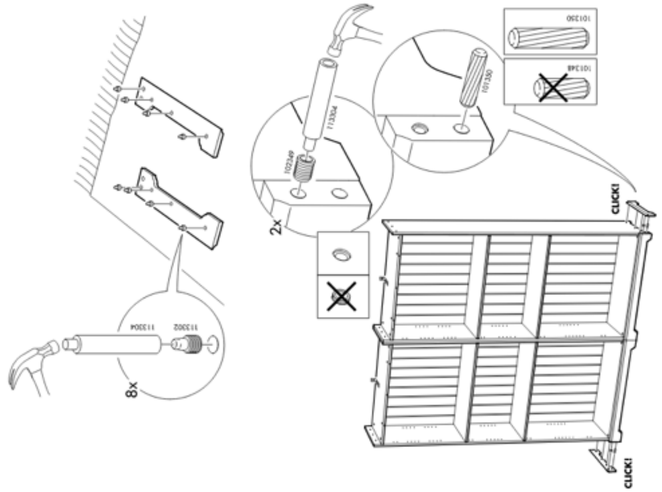
2. GET DÅTÅ



3. GET REFERENCE FILES

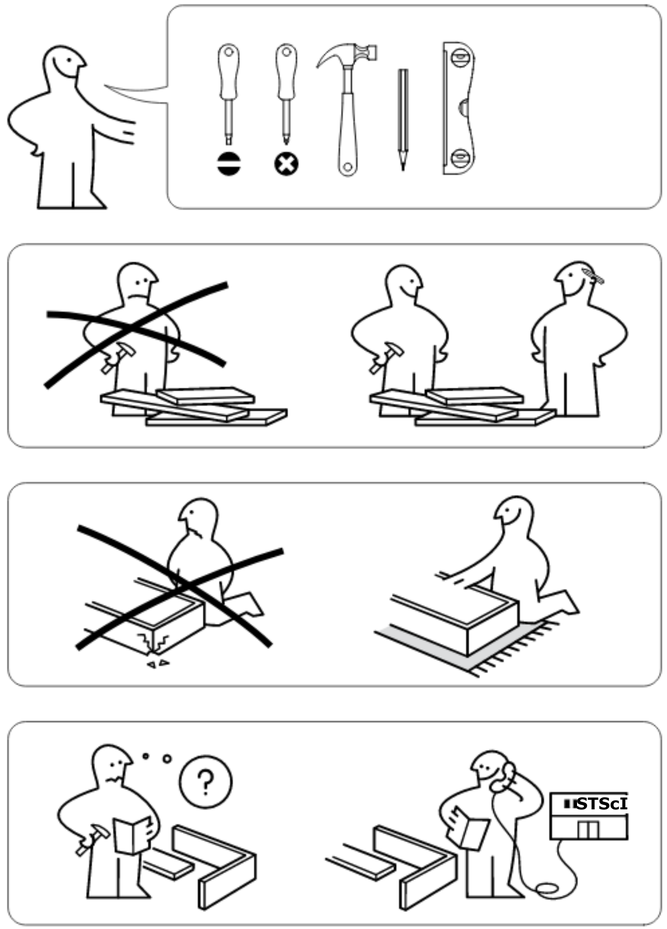


4. ASSEMBLE

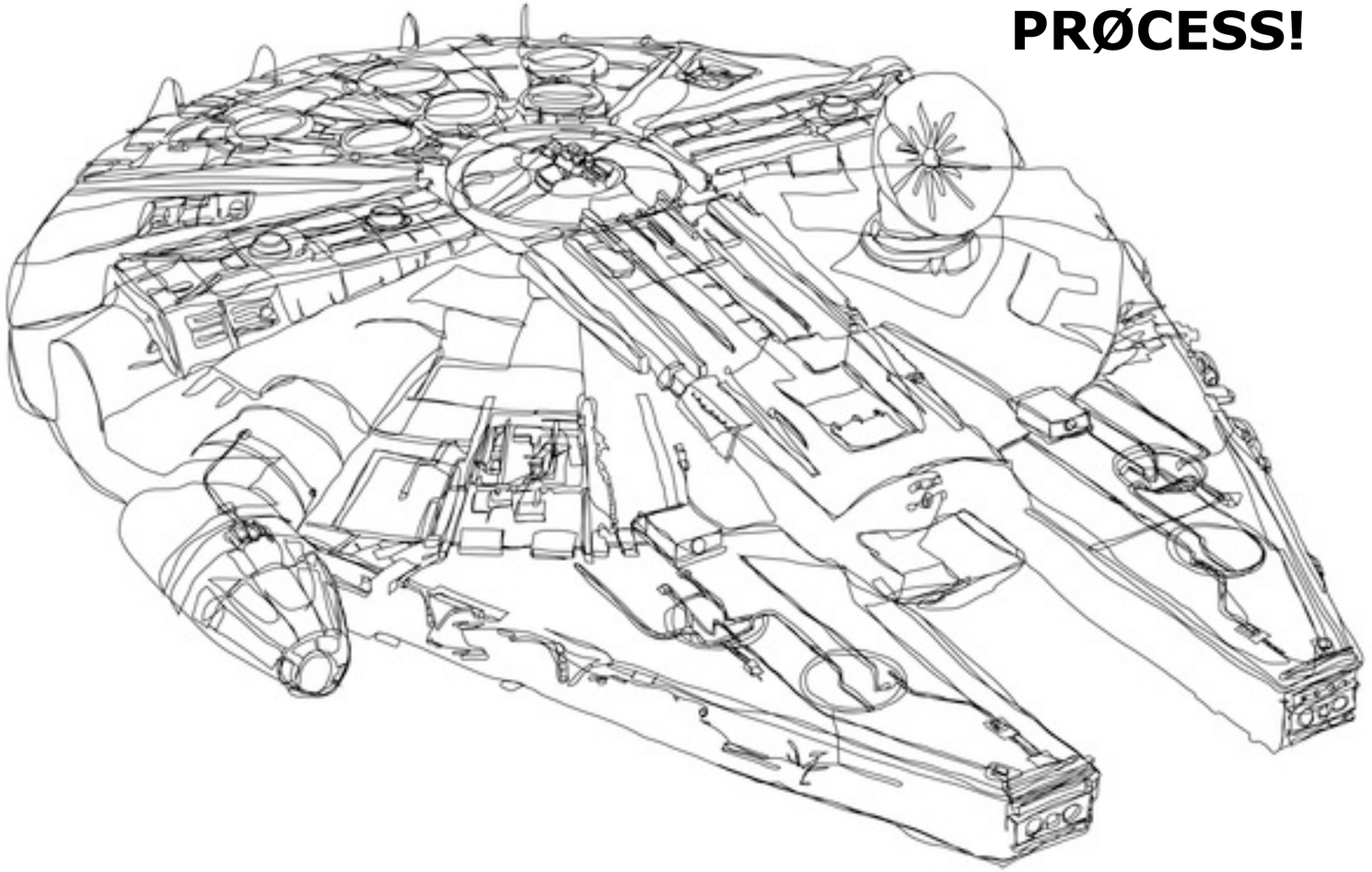


How Do We Process Data?

DÅTÅ PRØCESSING

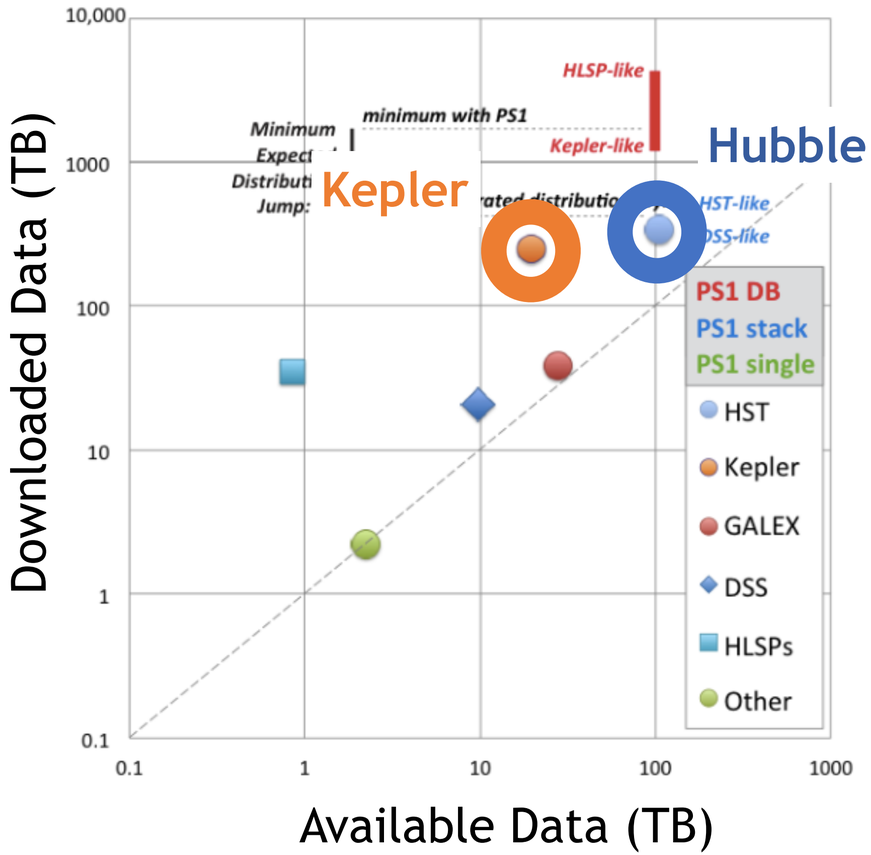


PRØCESS!

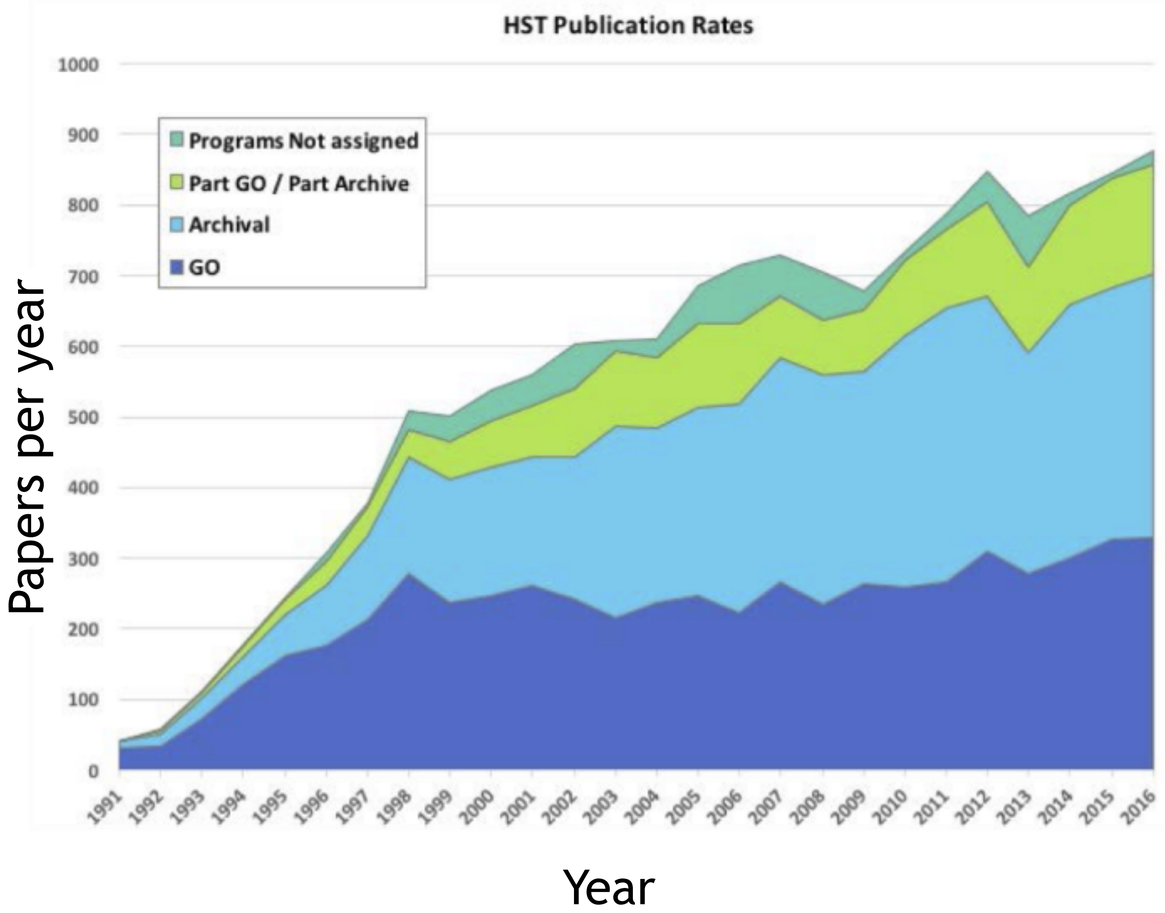


This Clearly Works

2011-2014



Big Data at STScI Report, 2016



Novacescu et al., 2016

This Clearly Works but ...

Will this work for JWST (and WFIRST)?

How can we lower the barrier to entry?

Is science use limited by capabilities?

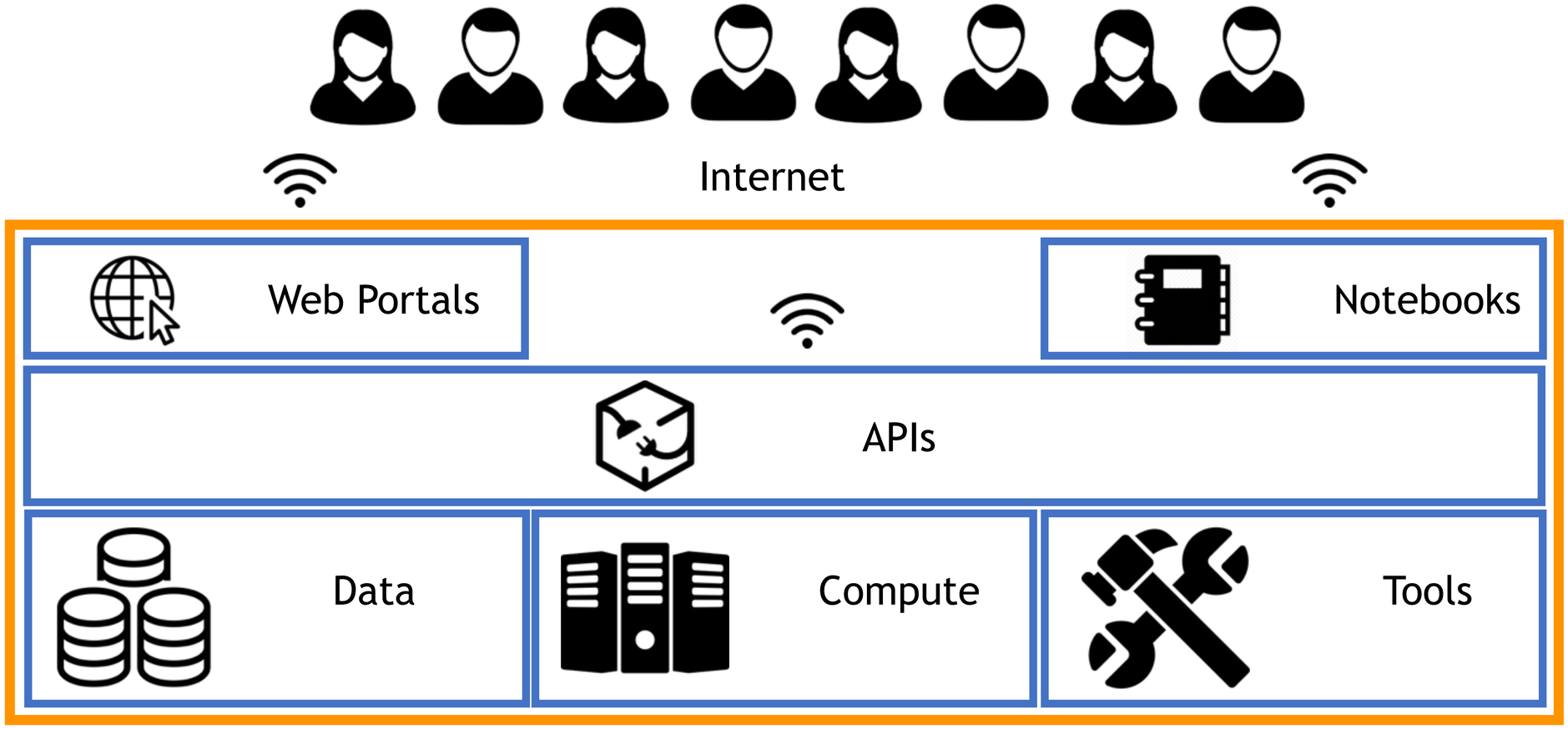
How do we improve provenance, repeatability, reproducibility?

Can we improve internal operations?

What is a Science Platform?

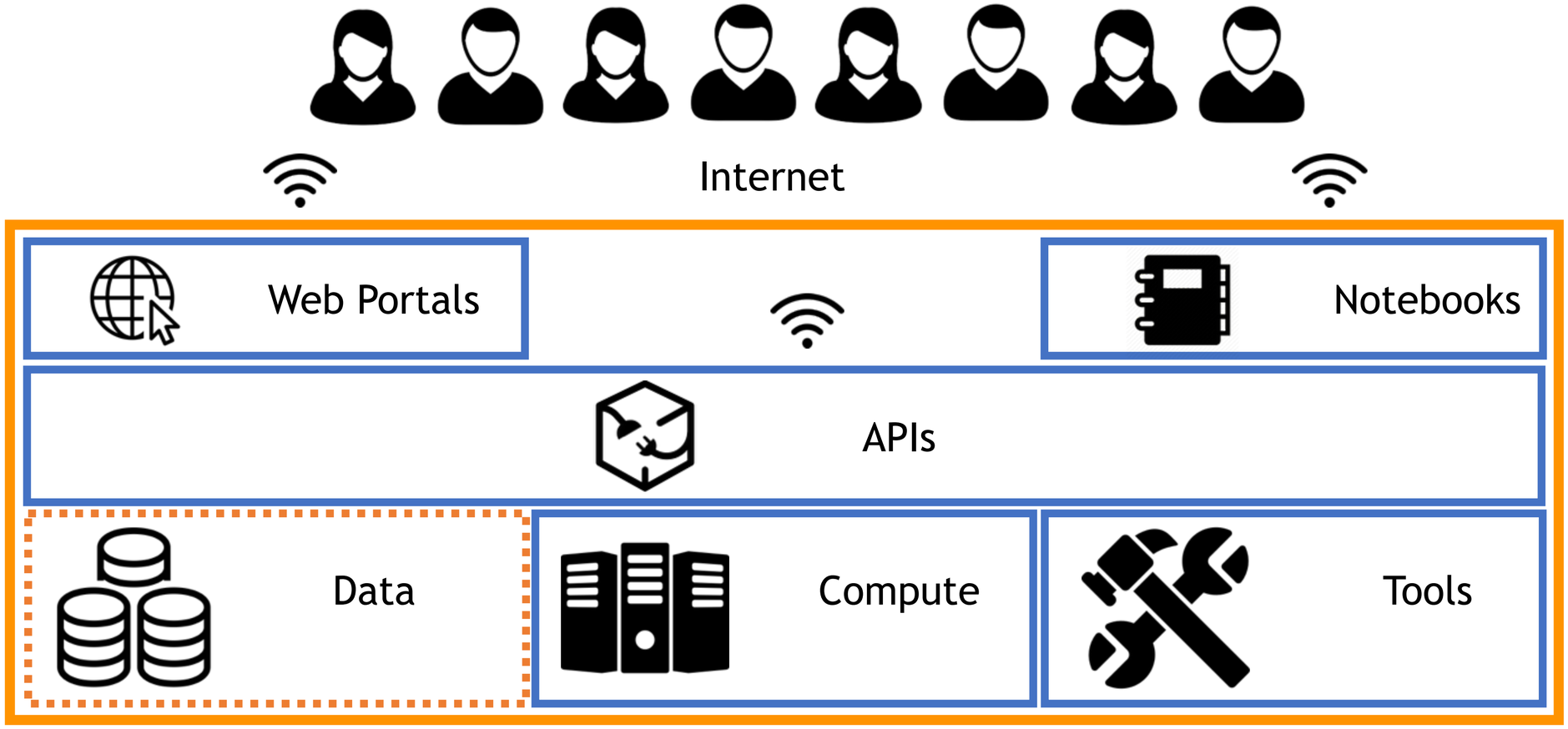
What is a Science Platform?

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



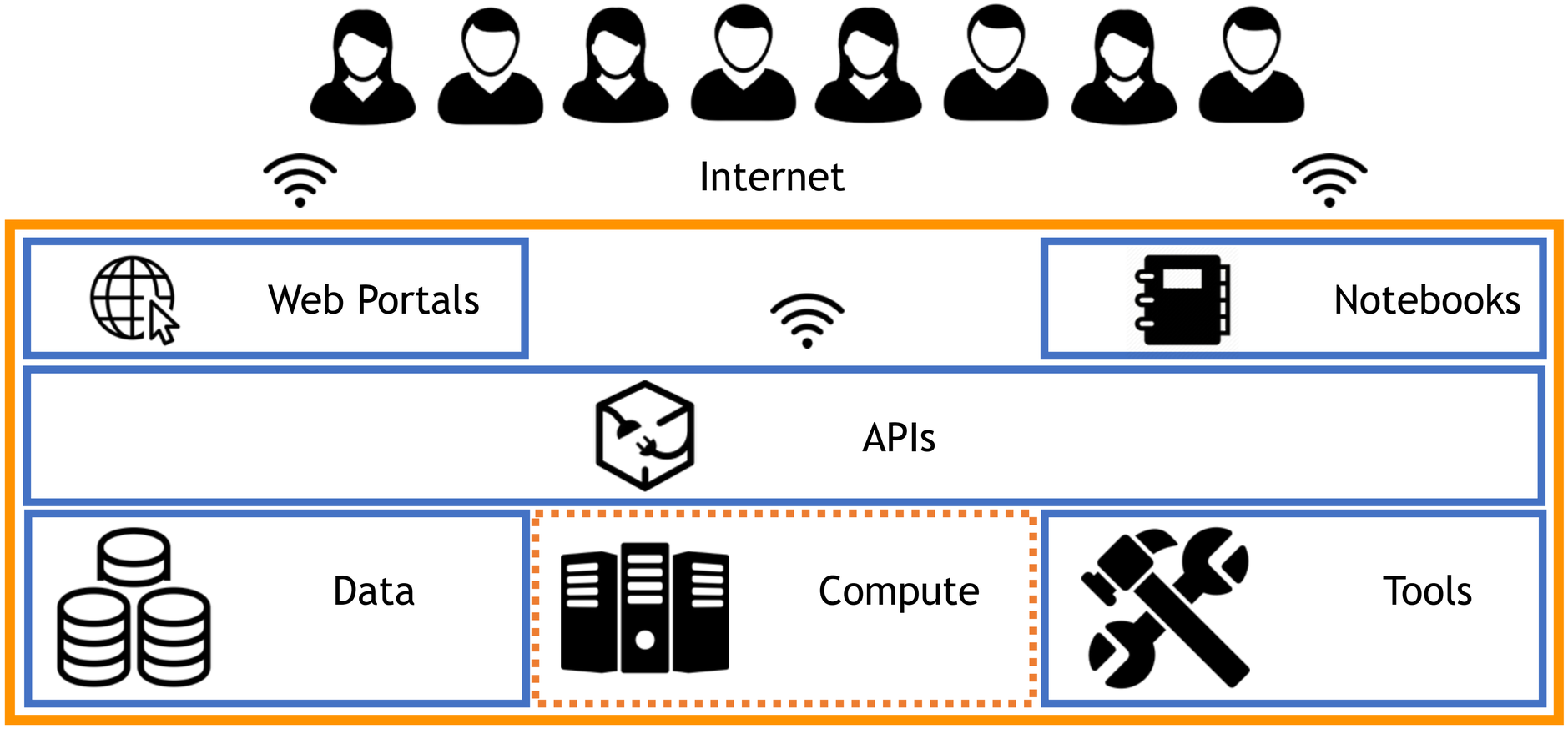
What is a Science Platform?

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



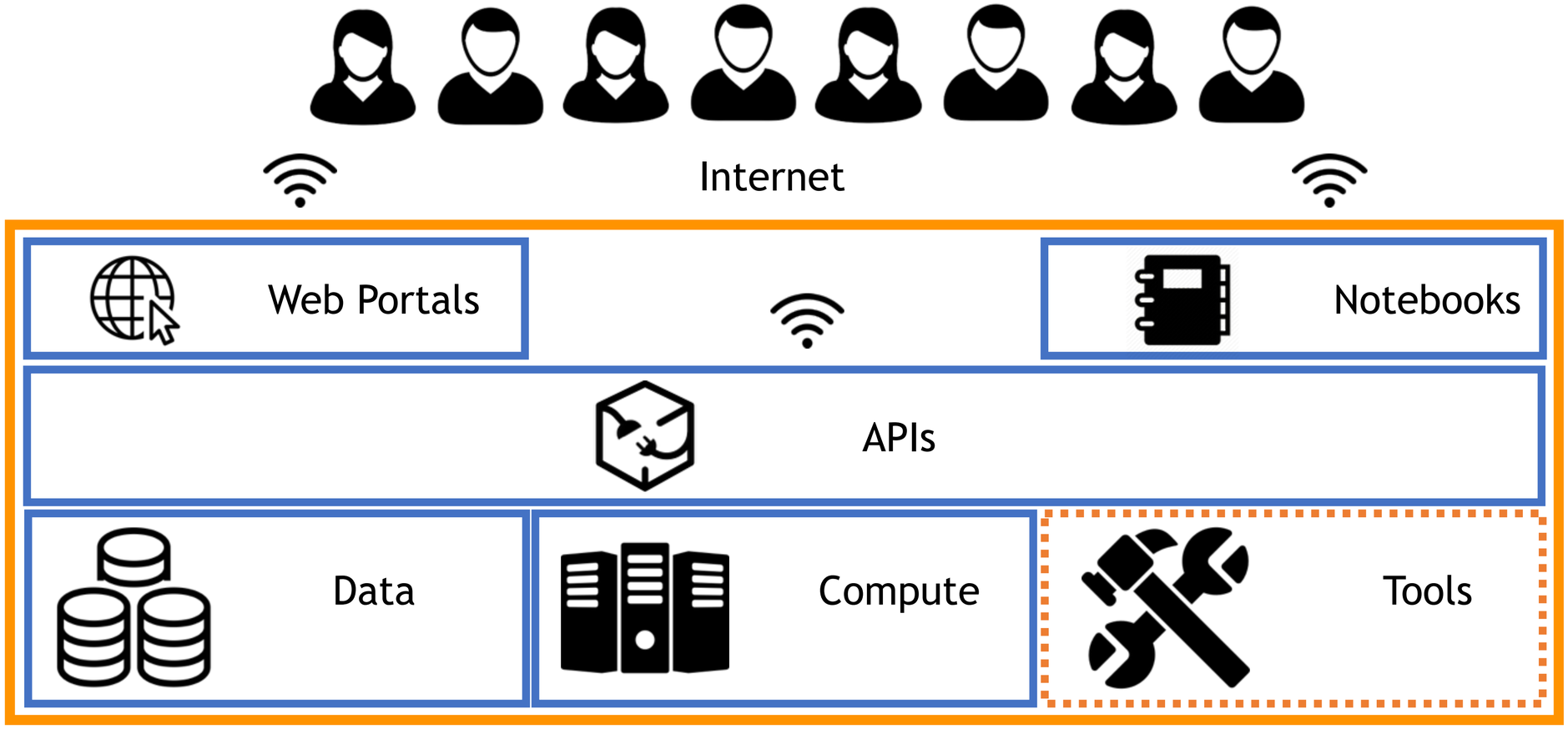
What is a Science Platform?

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



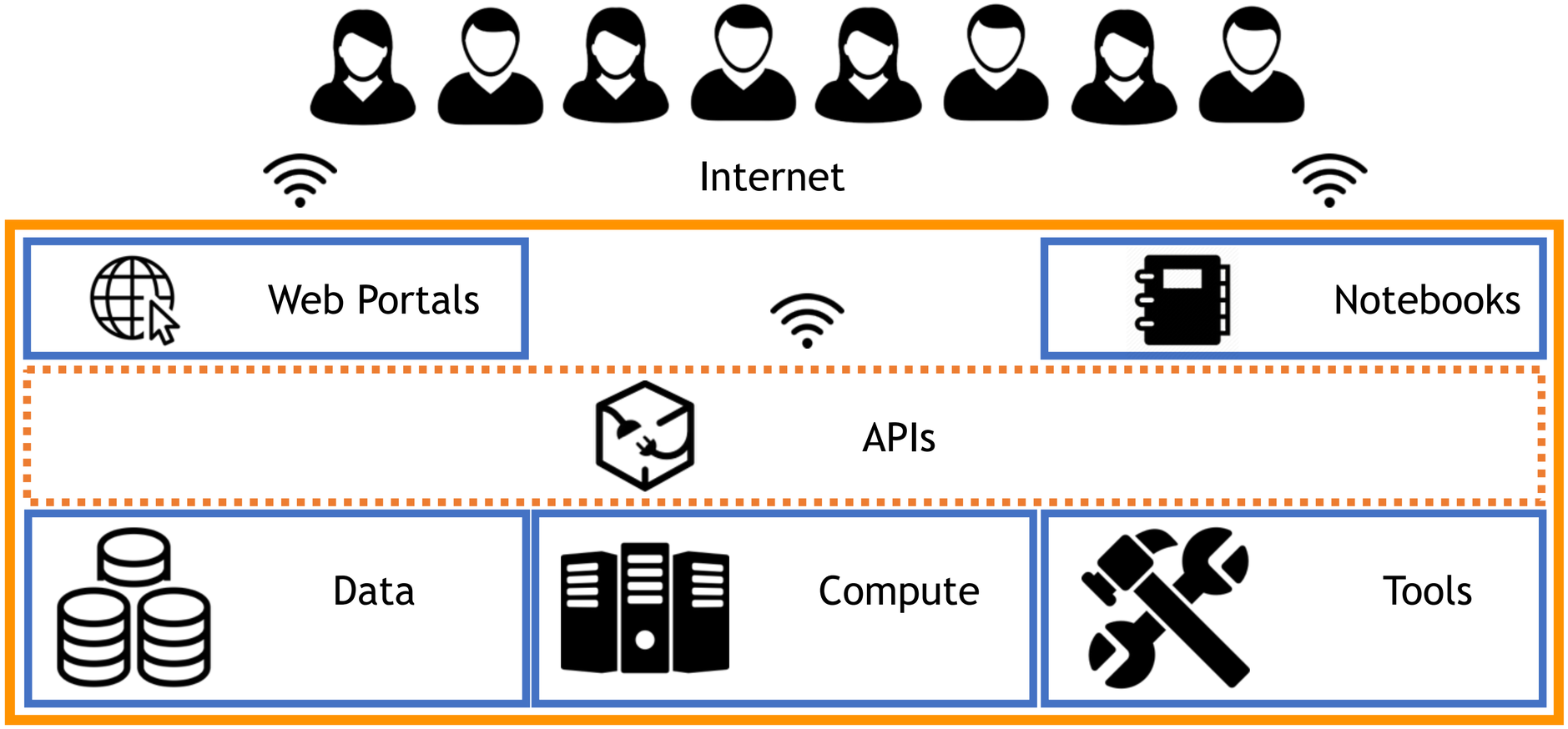
What is a Science Platform?

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



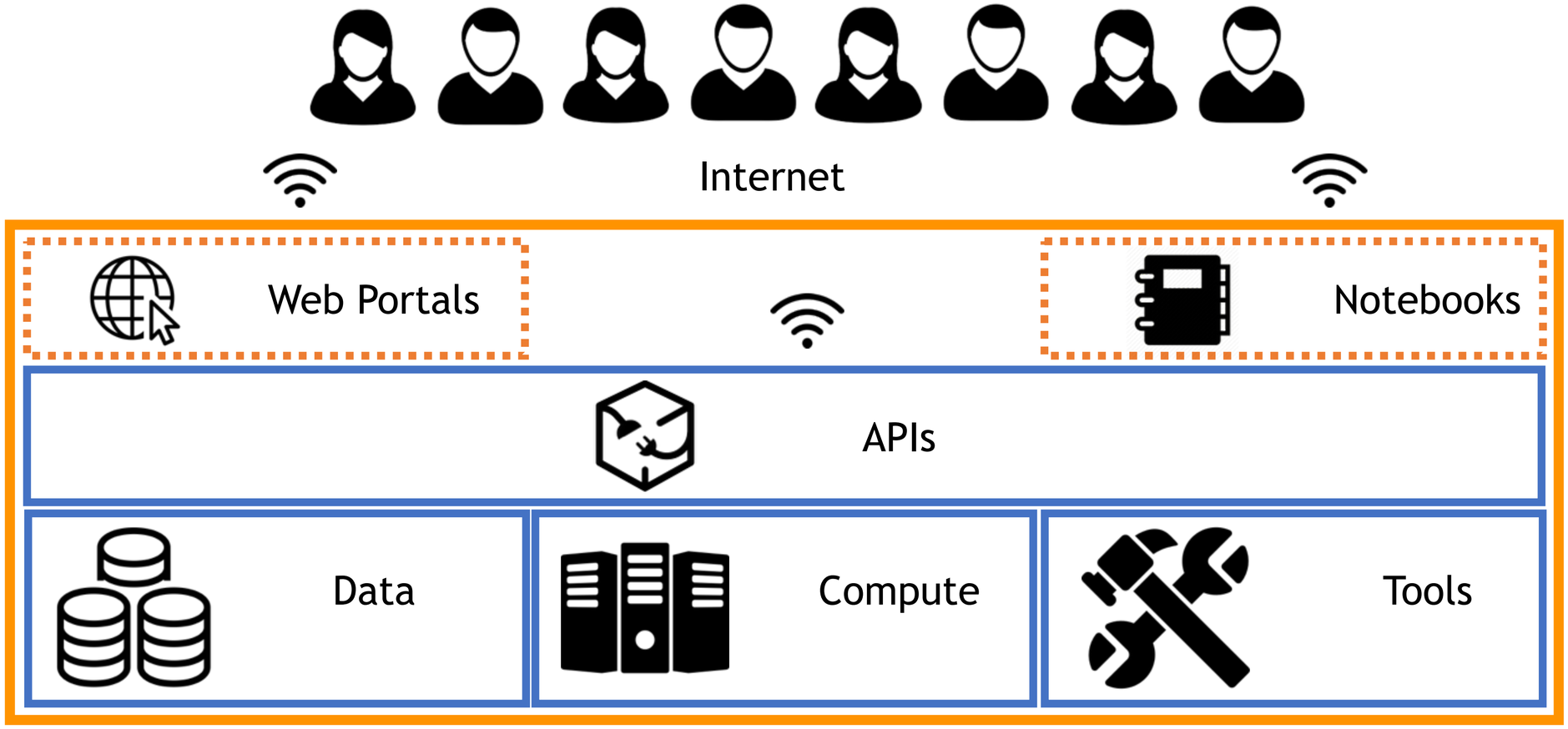
What is a Science Platform?

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



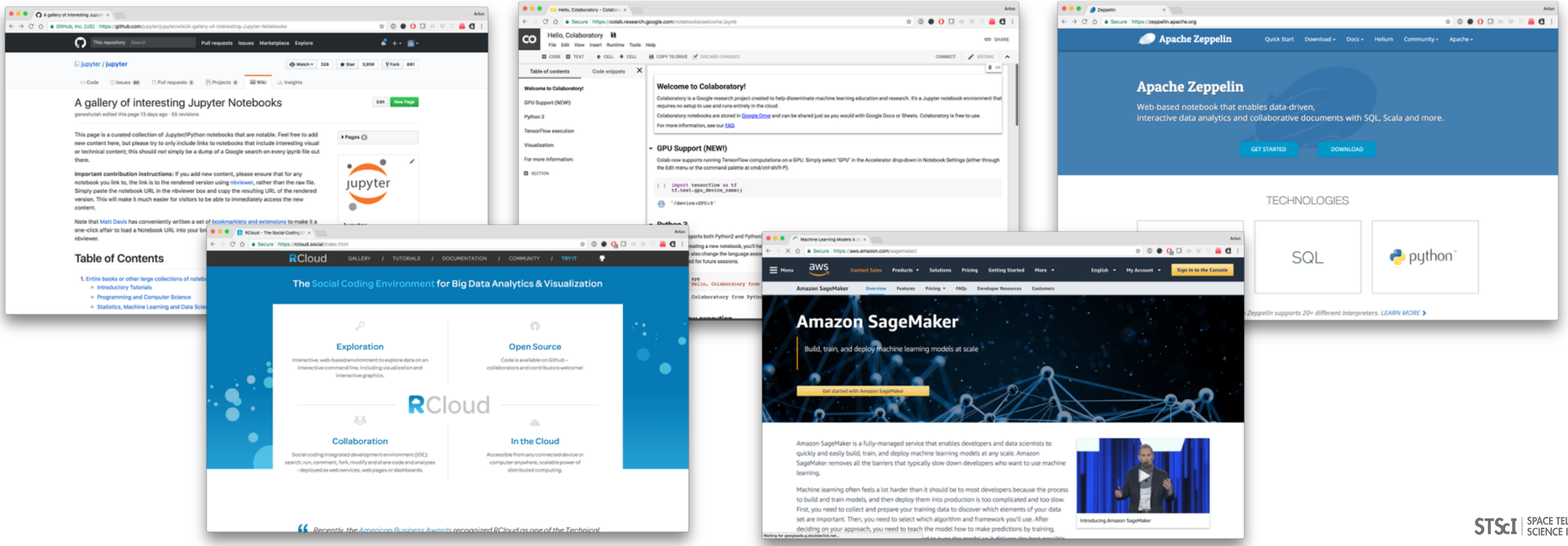
What is a Science Platform?

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



Technological Convergence

1. Notebook-driven analysis:
   - millions of Jupyter notebooks hosted on GitHub; RCloud; Apache Zeppelin; Google Colaboratory; AWS SageMaker, etc.



Technological Convergence

1. Notebook-driven analysis:

2. Compute is commodified by cloud providers

Technological Convergence

1. Notebook-driven analysis:

2. Compute is commodified by cloud providers

3. Software-defined infrastructure technologies are mature(ing)

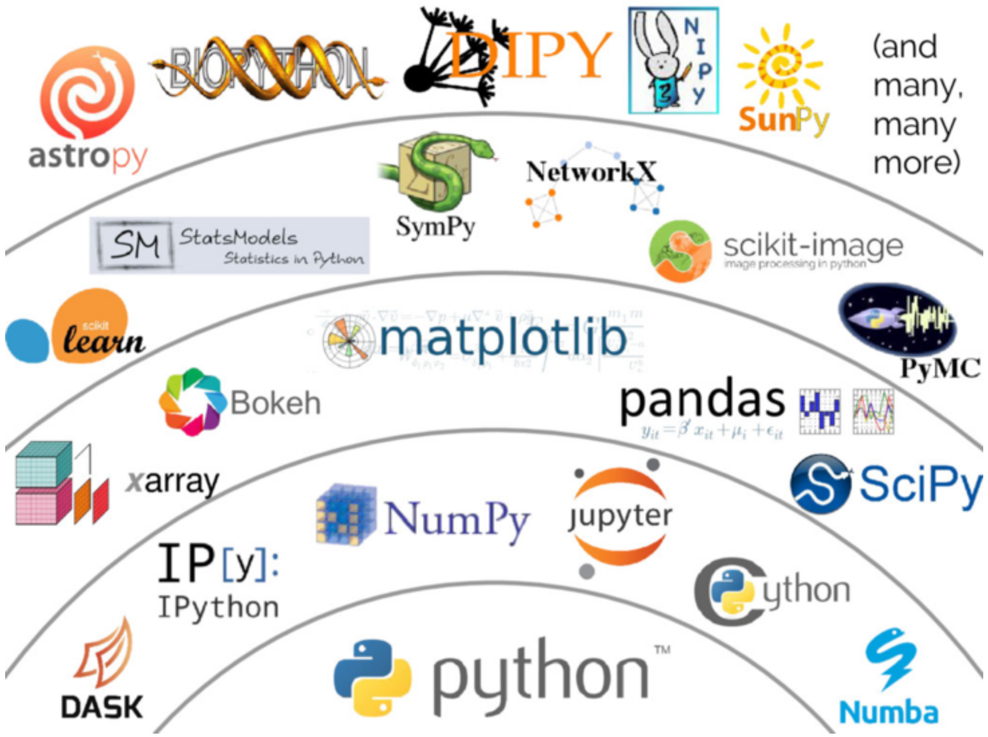
Technological Convergence

1. Notebook-driven analysis:

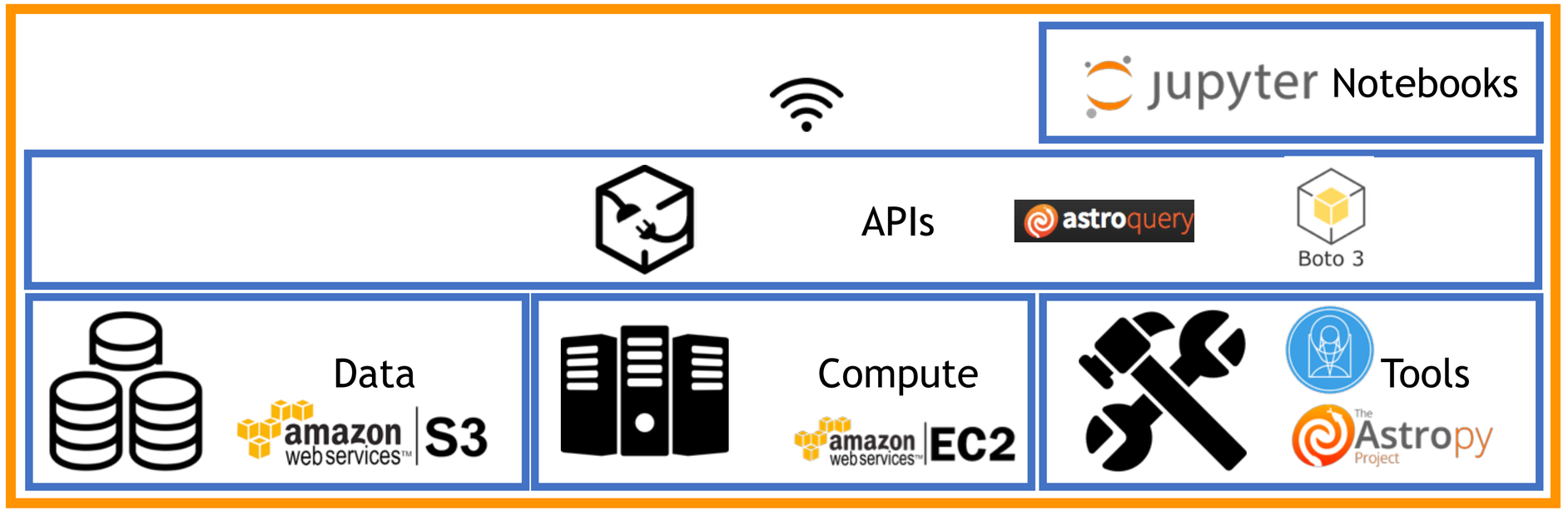
2. Compute is commodified by cloud providers

3. Software -defined infrastructure technologies are mature

4. A rich system of open source scientific compute



STScI Science Platform



STScI Science Platform

Registry of Open Data on AWS

Hubble Space Telescope Public Data

astronomy

Description

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.

Update Frequency

Hourly

License

STScI herby grants the non-exclusive, royalty free, non-transferable, worldwide right and license to use, reproduce and publicly display in all media public data from the Hubble Space Telescope.

Documentation

http://astroquery.readthedocs.io/en/latest/mast/mast.html

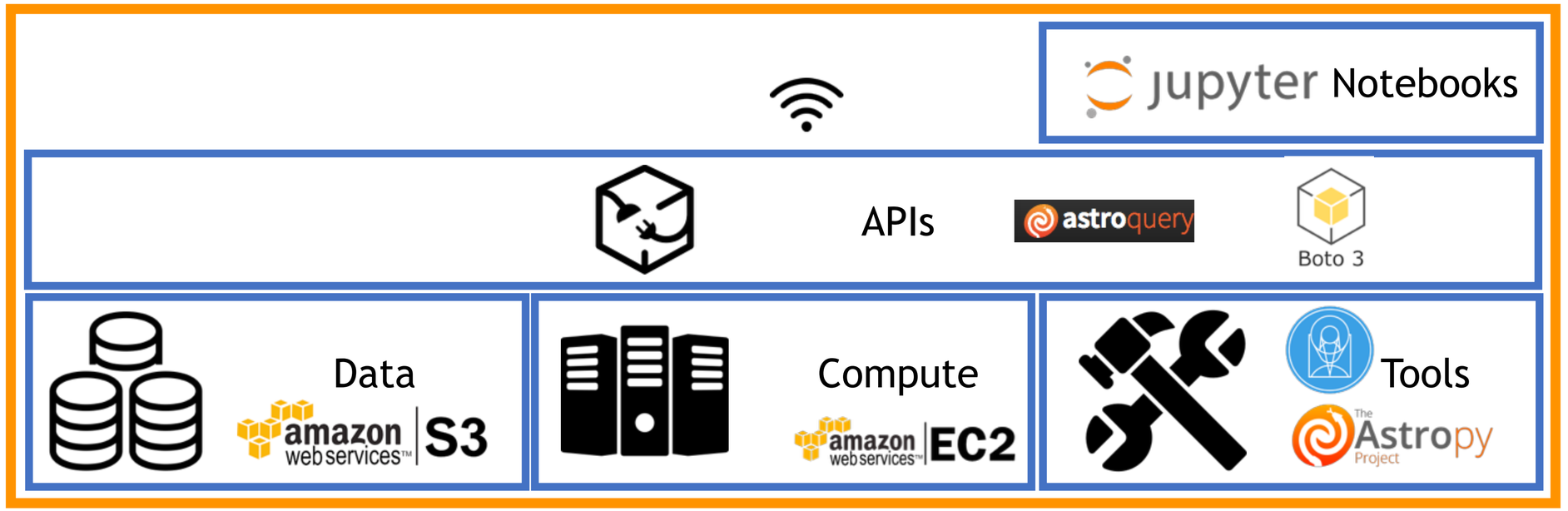
Contact

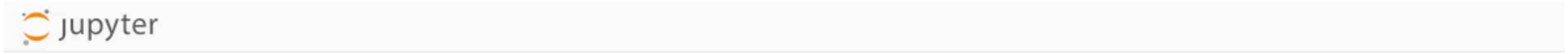
archive@stsci.edu

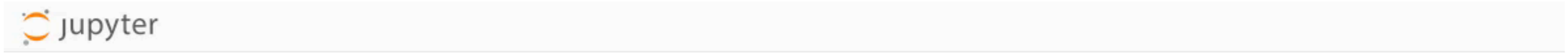
Usage Examples

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

STScI Science Platform

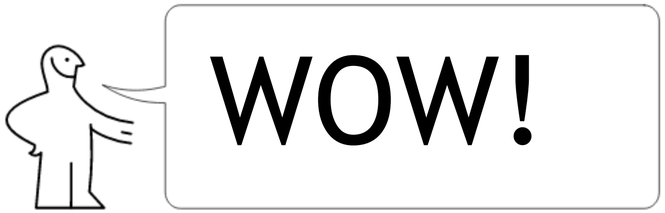



jupyter
Sign in with GitHub


jupyter
Sign in with GitHub

jupyter

Sign in with GitHub



Science Use Cases

- New archival research from small & large scale analysis of GO and survey data opened to a much broader community (lowering the barrier)
- Support new types of research: machine learning, deep learning, AI
- Analysis of simulated data in the same way as observations specifically when deployed at HPC centers
- Joint pixel level processing across different missions

Operations Use Cases

Running pipelines off-site

Collaboration between science and development teams (talk by R. Diaz)

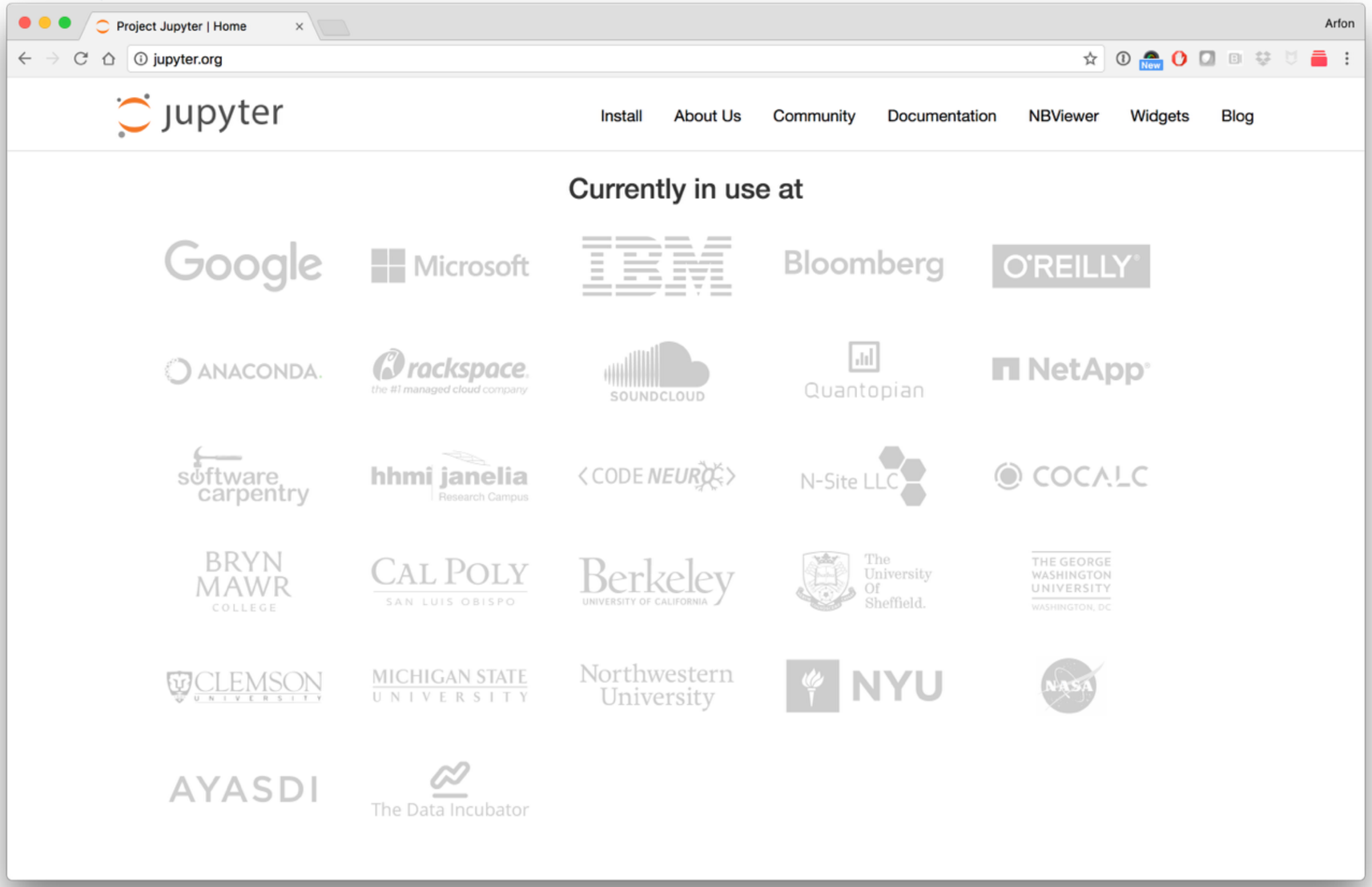
Large scale compute on archival data

ML for internal operations

How do you build a Science Platform?

Common technologies, many implementations

Not just in academia



Architecture

kubernetes

https://science.stsci.edu (SSO Auth)

STScI

API

API

docker docker docker ...

JupyterHub

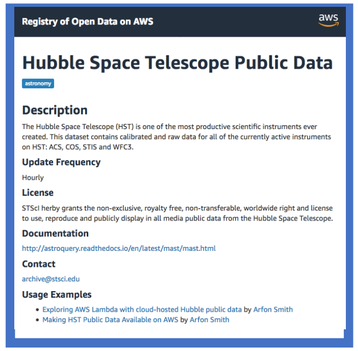
VM VM VM ...

Home Directories
e.g., /home/imomcheva

Docker container image registry

HST JWST

TESS ...



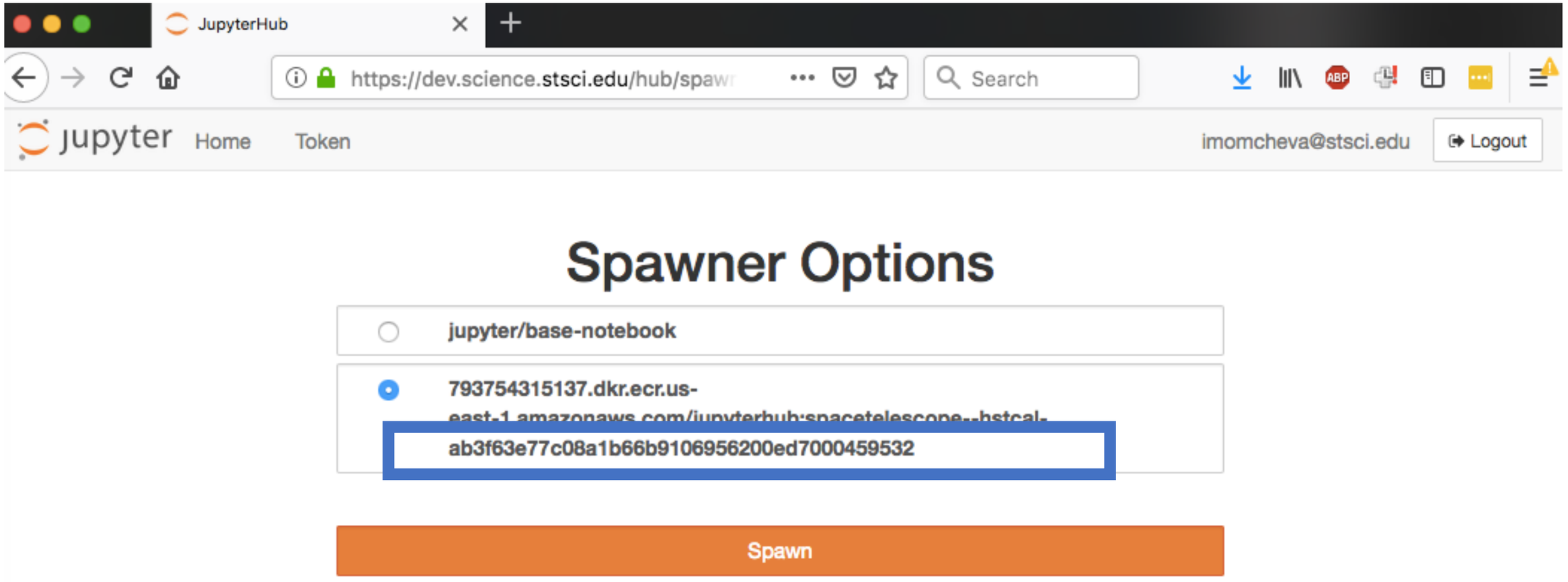
Docker (container images)

```
# Copyright (c) Association of Universities for Research in Astronomy
# Distributed under the terms of the Modified BSD License.
FROM jupyter/scipy-notebook

LABEL maintainer="Arfon Smith <arfon@stsci.edu>"

# Install Astroconda channel
RUN conda config --add channels http://ssb.stsci.edu/astroconda

# Create 'astroconda' channel configured with default packages
RUN conda create -n astroconda stsci python=3 -y

# Activate the astroconda channel
RUN ["/bin/bash", "-c", "source activate astroconda"]

# Install ipykernel switcher
RUN python -m ipykernel install --user \
    --name astroconda \
    --display-name "Python (astroconda)"



# Install ginga, ipywidgets and ipyevents for interactive plots
```

Docker (container images)

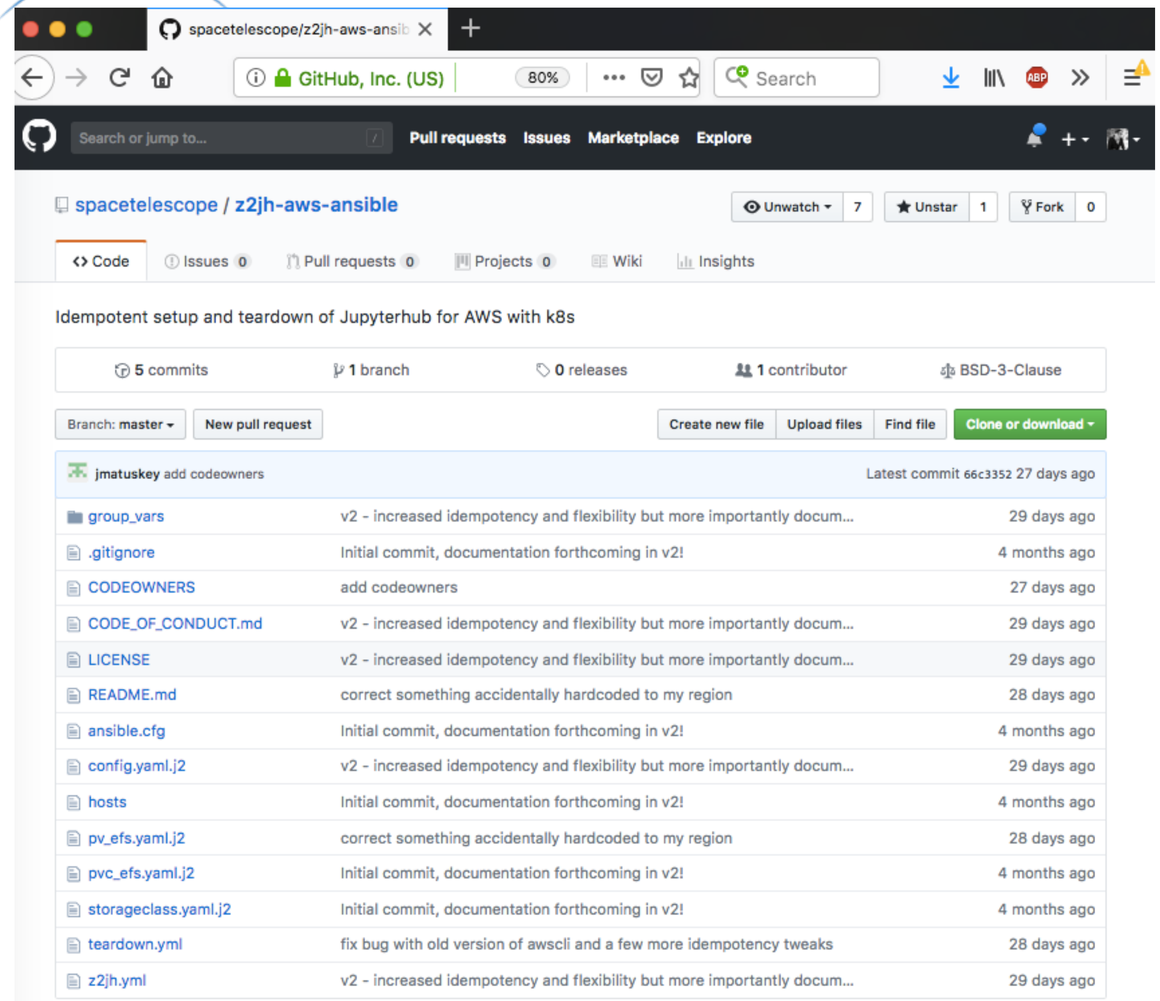
```
# Copyright (c) Association of Universities for Research in Astronomy
# Distributed under the terms of the Modified BSD License.
FROM jupyter/scipy-notebook

LABEL maintainer="Arfon Smith <arfon@stsci.edu>"

# Install Astroconda channel
RUN conda config --add channels http://ssb.stsci.edu/astroconda

# Create 'astroconda' channel configured with default packages
RUN conda create -n astroconda stsci python=3 -y

# Activate the astroconda channel
RUN ["/bin/bash", "-c", "source activate astroconda"]

# Install ipykernel switcher
RUN python -m ipykernel install --user \
    --name astroconda \
    --display-name "Python (astroconda)"



# Install ginga, ipywidgets and ipyevents for interactive plots
```

Composable machine images: FROM lsstsqre/pipeline


JupyterHub
https://dev.science.stsci.edu/hub/spawn
Search
jupyter
Home
Token
imomcheva@stsci.edu
Logout
Spawner Options
jupyter/base-notebook
793754315137.dkr.ecr.us-
ab3f63e77c08a1b66b9106956200ed7000459532
Spawn

Reference Deployment

https://github.com/spacetelescope/z2jh-aws-ansible



README.md

Zero to Jupyterhub for AWS in ansible

Ansible plays intended to set up a Jupyterhub instance from scratch. z2jh.yml tracks very closely with the AWS zero-to-jupyterhub readthedocs and idempotently sets up a Jupyterhub cluster. teardown.yml undoes up to a given level of the total installation, governed by which tags you specify. The default will only remove the Jupyterhub release.

Preconditions

- IAM role with attached policies: AmazonEC2FullAccess, IAMFullAccess, AmazonS3FullAccess, AmazonVPCFullAccess, AmazonElasticFileSystemFullAccess
- EC2 instance to serve as CI node provisioned (named [namespace]-ci) with key pair and above IAM role
- hosts file - put your CI node Public DNS (IPv4) as the only line of this
- group_vars/all
  - namespace - many things are named based on this for consistency
  - aws_region
  - ansible_ssh_private_key_file - absolute path of key file (.pem) which you use to ssh into the CI node
- Ansible installed on local machine

Zero To Jupyterhub play

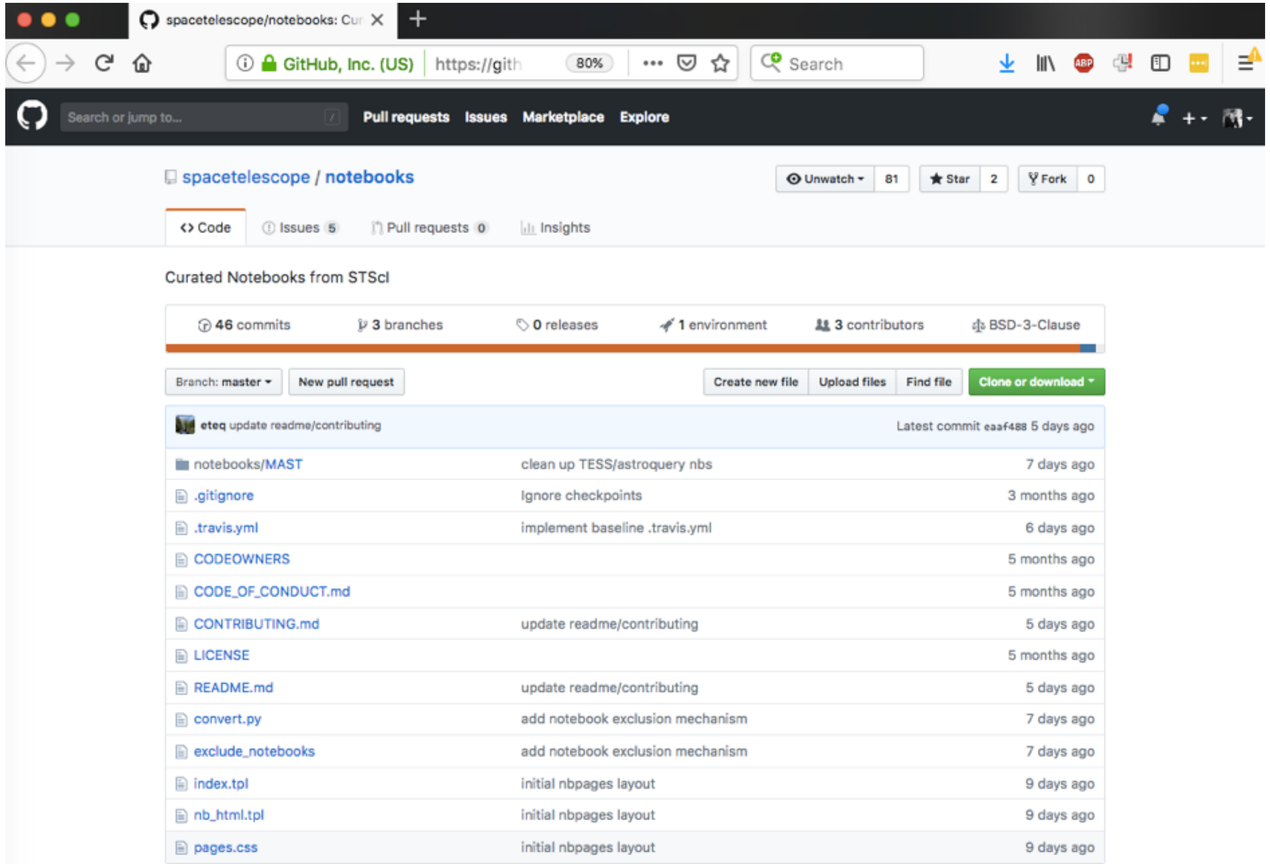
```
ansible-playbook -i hosts z2jh.yml -v
```

This will provision the AWS fixtures (EFS, S3) you need to create the infrastructure upon which Jupyterhub will run. It will create a Kubernetes cluster with kops as well and install Helm, Tiller, and download a given Jupyterhub chart and install it. Finally it will print the proxy URL where you navigate a browser to use your Jupyterhub.

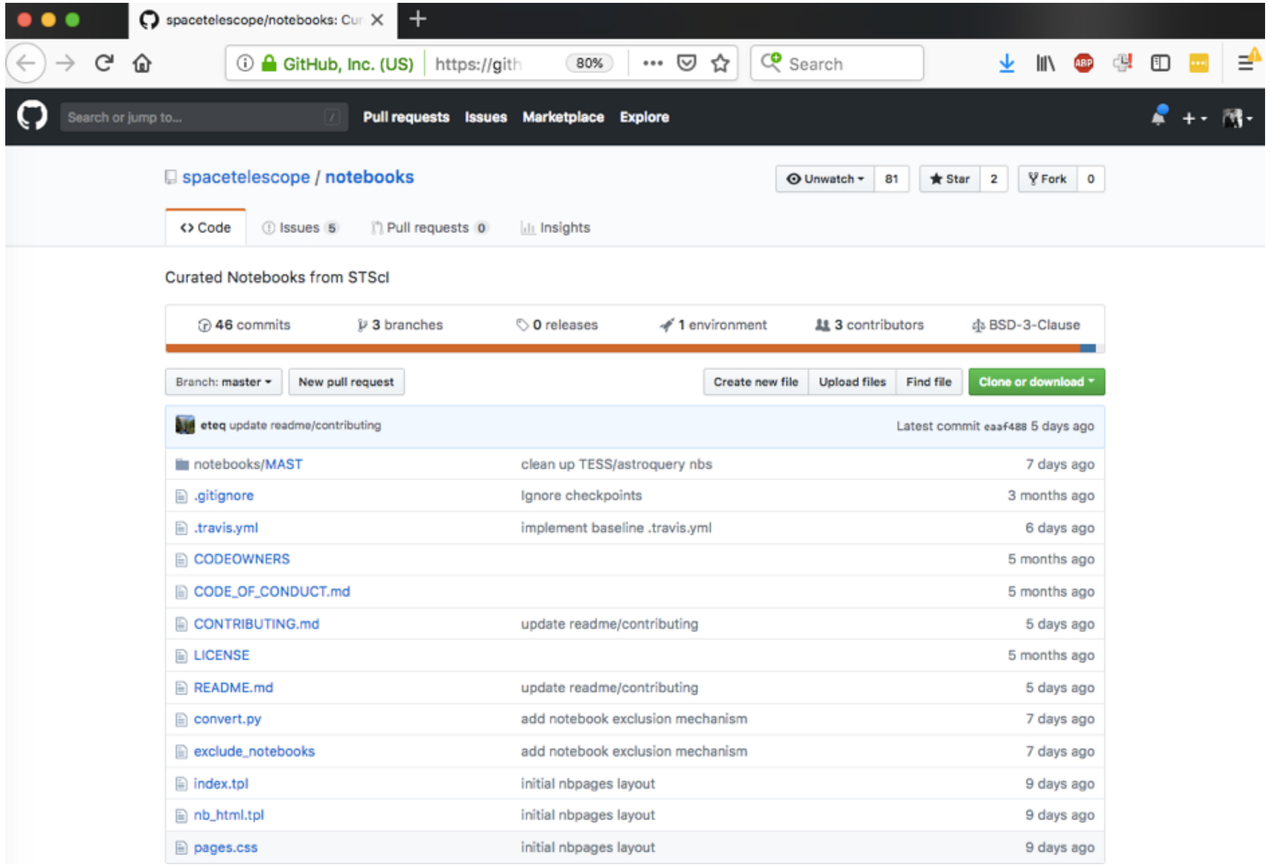
It is intended to be fully idempotent so feel free to run this and it will only create the fixtures and perform the operations if necessary. For example, if you already have an EFS called [namespace]-efs, it will not create a new one, it will use that. You could run it after manually deleting your Jupyterhub release and it would simply re-install a Jupyterhub release.

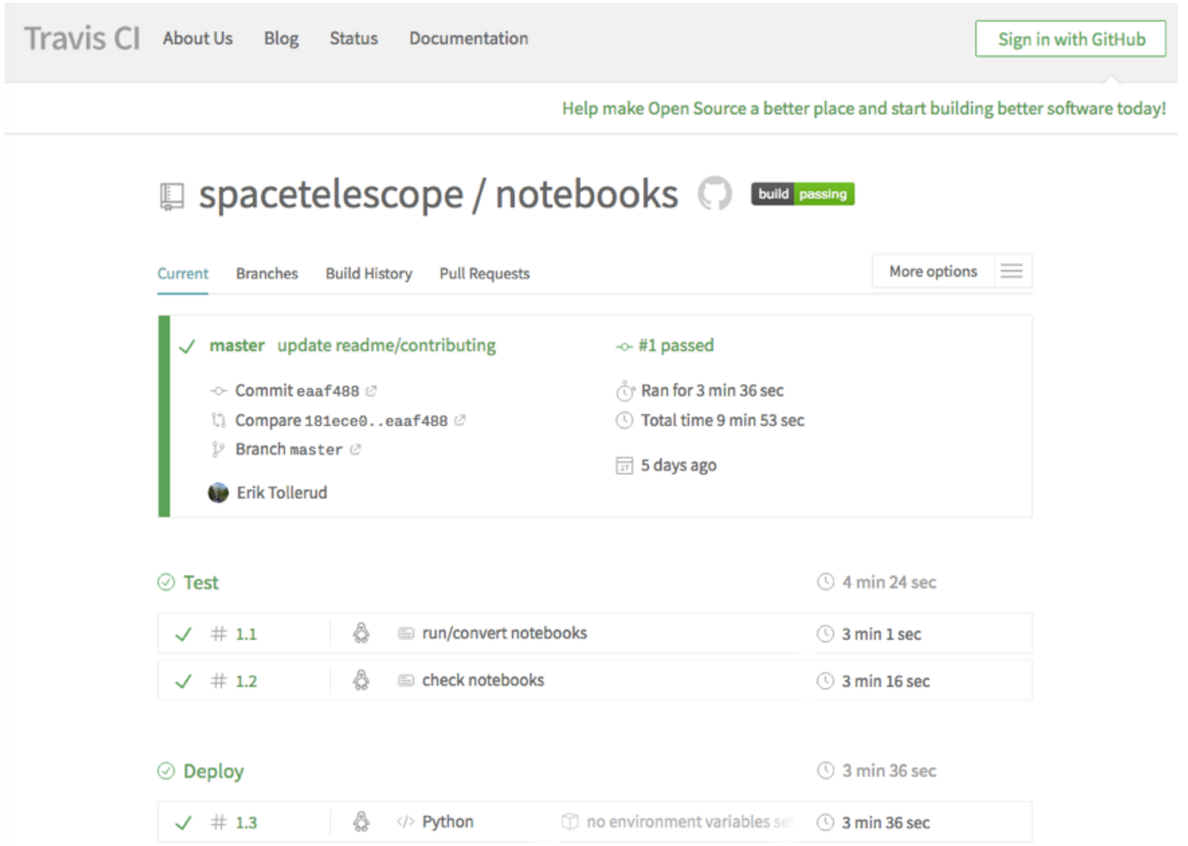
Modify the config templates as needed, these will generate the configs used in the helm install.

Content


spacetelescope/notebooks: Cur
GitHub, Inc. (US) https://gith 80%
Search
Search or jump to...
Pull requests Issues Marketplace Explore
spacetelescope / notebooks
Unwatch 81 Star 2 Fork 0
Code Issues 5 Pull requests 0 Insights
Curated Notebooks from STScI
46 commits 3 branches 0 releases 1 environment 3 contributors BSD-3-Clause
Branch: master New pull request
Create new file Upload files Find file Clone or download
eteq update readme/contributing Latest commit eaaf488 5 days ago
notebooks/MAST clean up TESS/astroquery nbs 7 days ago
.gitignore Ignore checkpoints 3 months ago
.travis.yml implement baseline .travis.yml 6 days ago
CODEOWNERS 5 months ago
CODE_OF_CONDUCT.md 5 months ago
CONTRIBUTING.md update readme/contributing 5 days ago
LICENSE 5 months ago
README.md update readme/contributing 5 days ago
convert.py add notebook exclusion mechanism 7 days ago
exclude_notebooks add notebook exclusion mechanism 7 days ago
index.tpl initial nbpages layout 9 days ago
nb_html.tpl initial nbpages layout 9 days ago
pages.css initial nbpages layout 9 days ago

Content


spacetelescope/notebooks: Cur
GitHub, Inc. (US) https://gith 80%
Search
Search or jump to...
Pull requests Issues Marketplace Explore
spacetelescope / notebooks
Unwatch 81 Star 2 Fork 0
Code Issues 5 Pull requests 0 Insights
Curated Notebooks from STScI
46 commits 3 branches 0 releases 1 environment 3 contributors BSD-3-Clause
Branch: master New pull request
Create new file Upload files Find file Clone or download
eteq update readme/contributing Latest commit eaaf488 5 days ago
notebooks/MAST clean up TESS/astroquery nbs 7 days ago
.gitignore Ignore checkpoints 3 months ago
.travis.yml implement baseline .travis.yml 6 days ago
CODEOWNERS 5 months ago
CODE_OF_CONDUCT.md 5 months ago
CONTRIBUTING.md update readme/contributing 5 days ago
LICENSE 5 months ago
README.md update readme/contributing 5 days ago
convert.py add notebook exclusion mechanism 7 days ago
exclude_notebooks add notebook exclusion mechanism 7 days ago
index.tpl initial nbpages layout 9 days ago
nb_html.tpl initial nbpages layout 9 days ago
pages.css initial nbpages layout 9 days ago


Travis CI About Us Blog Status Documentation
Sign in with GitHub
Help make Open Source a better place and start building better software today!
spacetelescope / notebooks
build passing
Current Branches Build History Pull Requests
More options
master update readme/contributing
#1 passed
Commit eaaf488
Compare 181ece0..eaaf488
Branch master
Ran for 3 min 36 sec
Total time 9 min 53 sec
5 days ago
Erik Tollerud
Test
4 min 24 sec
1.1 run/convert notebooks 3 min 1 sec
1.2 check notebooks 3 min 16 sec
Deploy
3 min 36 sec
1.3 Python 3 min 36 sec

Challenges and Future Directions

Challenges

- Wide variety of use cases: need flexibility in machine types & containers
- “Lifting & shifting” workflows not possible: need more better docs, more notebooks
- Notebooks can be problematic: hidden states
- Collaborative workflows currently not possible: dev is happening
- Billing model & user quotas: deploy your own?
- User management: privacy vs. security
- Running batch compute

Convergence of Approaches?

Birds of a Feather session on Science platforms

William O'Mullane[1], Megan Sosey[2], Hassan Siddiqui[4], Gregory Dubois-Felsmann[3], Gerard Lemson[5], Christophe Arviset[6], Mike Fitzpatrick[7], Ivelina Momcheva[2], Sebastien Fabbro[8], Brian Major[8]
[1]*Large Synoptic Survey Telescope, Tucson, AZ, USA;* womullan@lsst.org

[2]*Space Telescope Science Institute*

[3]*IPAC, California Institute of Technology, Pasadena, CA, U.S.A.*

[4]*Vega for Gaia/ESAC*

[5]*The Johns Hopkins University*

[6]*European Space Astronomy Centre*

[7]*NOAO*

[8]*CADC*

Abstract. How users will interact with data in the future is always unclear. Currently we see Jupyter Notebooks or JupyterLab emerging in many places as the way forward for one aspect of this. This BoF explored some topics around providing and environment for doing science.

O'Mullane et al., 2017 ADASS Proceedings

https://github.com/spacetelescope/science-platforms-workshop

Summary

- **Science Platforms allow users to run analysis next to data, real or simulated**
- **A solution to Big Data in astro**
- **But also key for for science applications NOW:**
  - Small data users benefit as well
  - Include tutorial notebooks to ramp up users
  - Remove software installation as a barrier to entry
  - Provide access to a range of computational resources
  - Allows for easier reproducibility
- **And for internal operations:**
  - Improve development cycle
  - Expand capabilities

Science Platforms are possible today with existing off-the-shelf technologies.

“The future is here now - it’s just not very evenly distributed”

William Gibson

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Thank you!

Ivelina Momcheva

imomcheva@stsci.edu