Scientific Platforms and Astronomical Data Access in the Era of (Scientific) Cloud Computing

Matias Carrasco Kind
Senior Research Scientist, NCSA/UIUC
Data Release Scientist, DES

AstroData2020s Science Workshop, Caltech
Dec 4th, 2018

Outline

- What does Data Access mean?
- Scientific Platforms and Gateways
- The Notebook revolution
- Scientific Cloud computing
- Containerization
- Kubernetes
- Applications

What is a Data Release?

Data Products

Interfaces

Documentation

Support

What is a Data Release?

Data Products

- Preparation
- Vetting
- Versioning
- Consistency
- Integrity
- Redundancy
- Data Model
- Storage
- Backups
- Recovery
- Hardware

Interfaces

- Development
- Version control
- Licenses
- Data Access
- Languages
- Sustainability
- Guidelines
- Scalability
- Deployment
- Hardware
- Maintenance

Documentation

- Papers
- Web
- Code
- Data Model
- Data Access
- Data Format
- Guidelines
- Accessible
- Maintenance
- Contributions

Support

- Short Term
- Long Term
- Forum
- Help
- Understanding
- Deployment
- Privacy
- Maintenance
- Focused
- Distributed

What is a Data Release?

Data Products

- Preparation
- Vetting
- Versioning
- Consistency
- Integrity
- Redundancy
- Data Model
- Storage
- Backups
- Recovery
- Hardware

Interfaces

- Development
- Version control
- Licenses
- Data Access
- Languages
- Sustainability
- Guidelines
- Scalability
- Deployment
- Hardware
- Maintenance

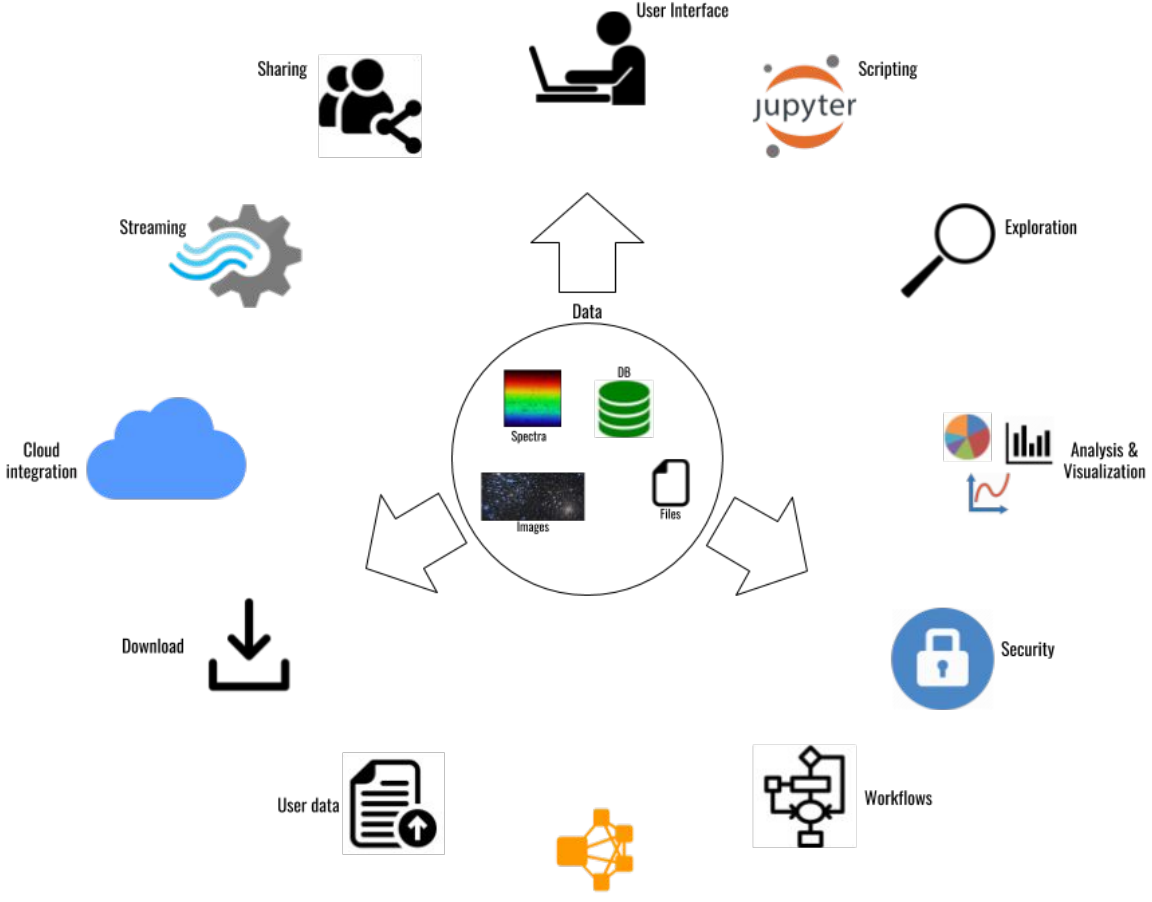
Documentation

- Papers
- Web
- Code
- Data Model
- Data Access
- Data Format
- Guidelines
- Accessible
- Maintenance
- Contributions

Support

- Short Term
- Long Term
- Forum
- Help
- Understanding
- Deployment
- Privacy
- Maintenance
- Focused
- Distributed

What is Data Access?

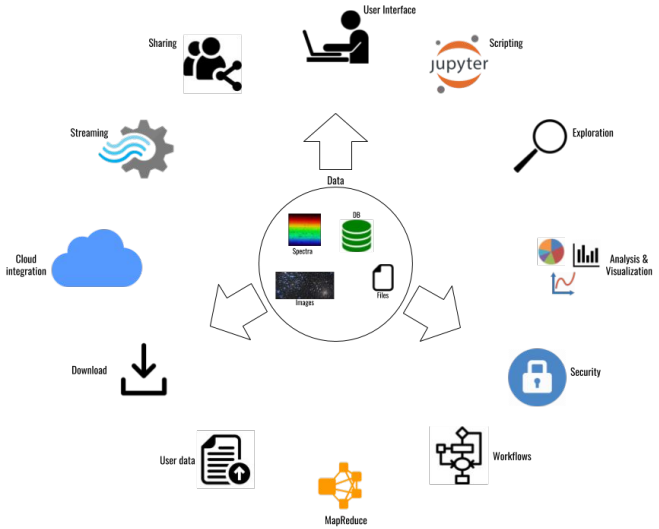


Several meanings around a central data archive, a.k.a “data lake”, repository with common components

- Storage
- Security
- Retrieving
- Interacting
- Modifying
- Understanding

Scientific Platforms and Gateways

... and many of these concepts are also associated with Scientific Platforms and Gateways (and Science portals, Science servers, etc.)



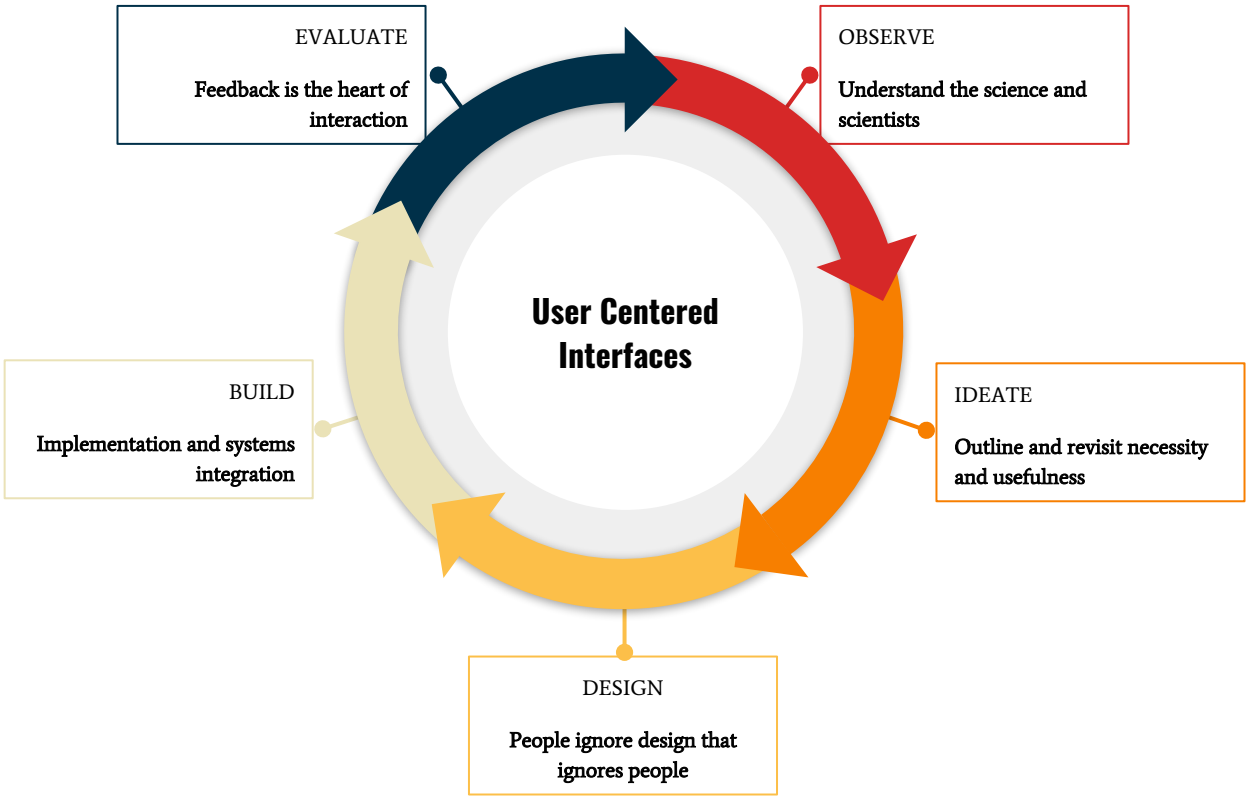
"Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines." (Science Gateways Institute)

"Science gateways is a place to do collaborative scientific related activities" (Me)

User (Scientist) Centered Design

Data Access would not exists without a user interface, but will only succeed if it is user driven.

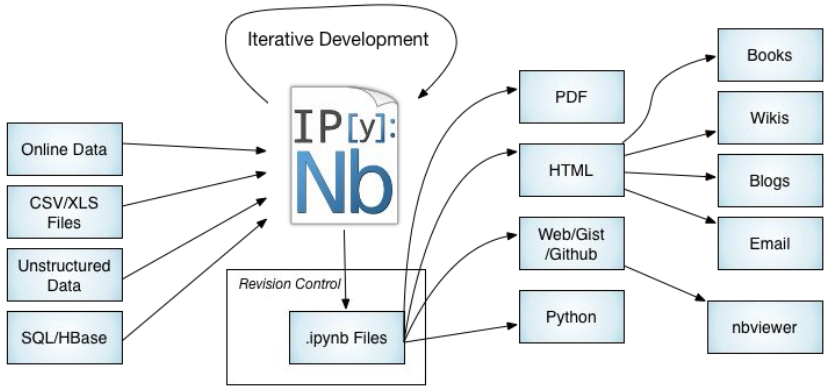
"... In an ideal world, a user would remember every function after only a single use, but we do not live in idealism. The reality is that familiarity and intuition must be consciously designed into the interface"

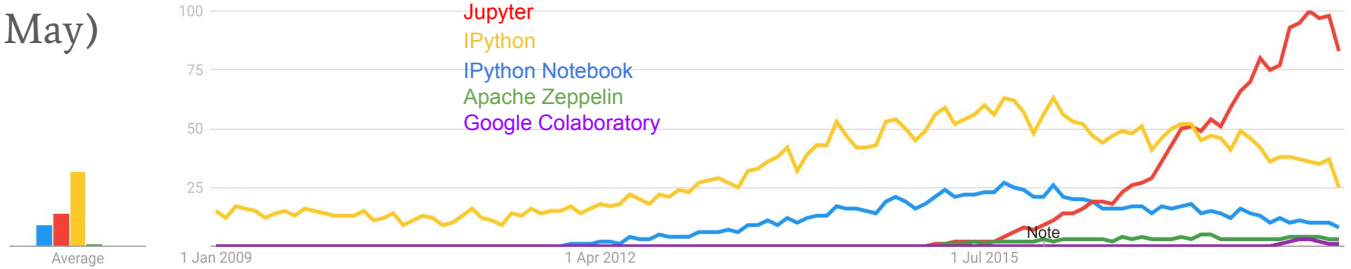


The Notebook Revolution

The Notebook Development

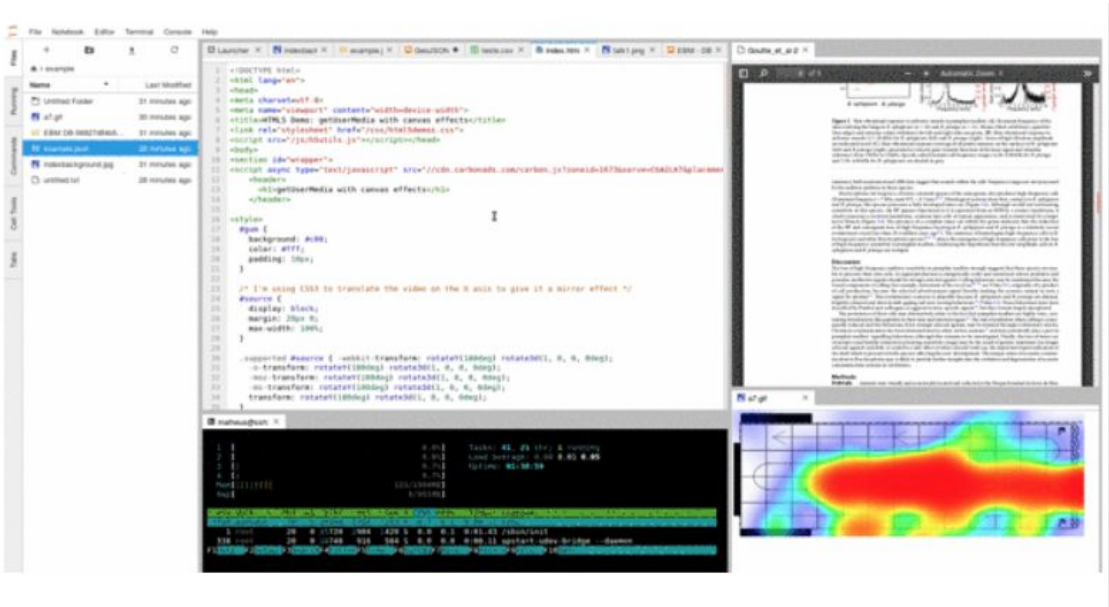
- Started from ideas like Matlab, Maple or Mathematica ~1988
- IPython has been around since 2001
- Sage Notebook released in 2005 (uses IPython)
- IPython Notebook was released in 2011
- IPython Notebook moved to Jupyter in 2014
- Apache Zeppelin created in 2015 (JVM and integrated with Apache Products)
- Beaker Notebook 2015 (moved to BeakerX)
- Google Colaboratory released in Oct 2017 (from ideas back in 2014)
- Cocalc (by SageMath) in 2018
- Jupyter Lab Beta 2.0 (May)





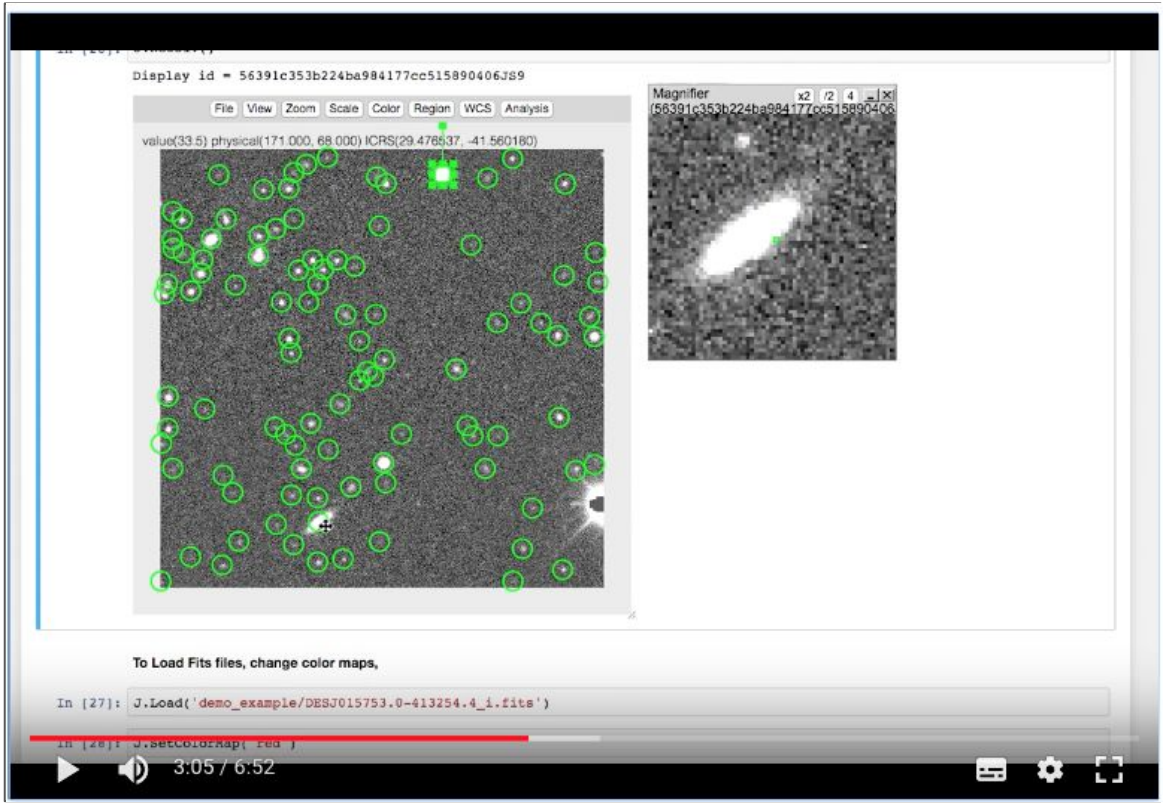
The Jupyter Notebook

- Computational narrative
- Scripting interface
- Scientific oriented interface
- Customizable
- Collaborative
- Adopted by many projects in scientific fields
- Widgets
- Big Data Integration (Spark
- Interactive plots
- Multiple Kernels (Python, R, Julia, Scala, etc.)



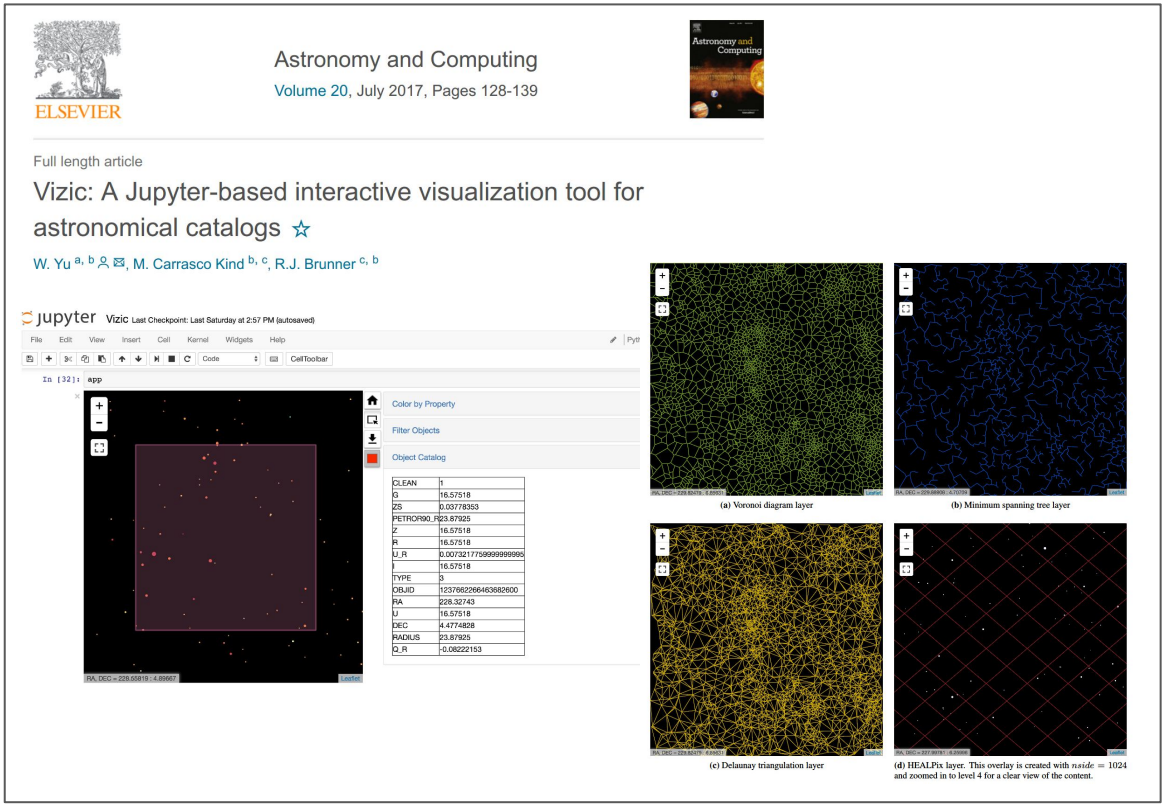
Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



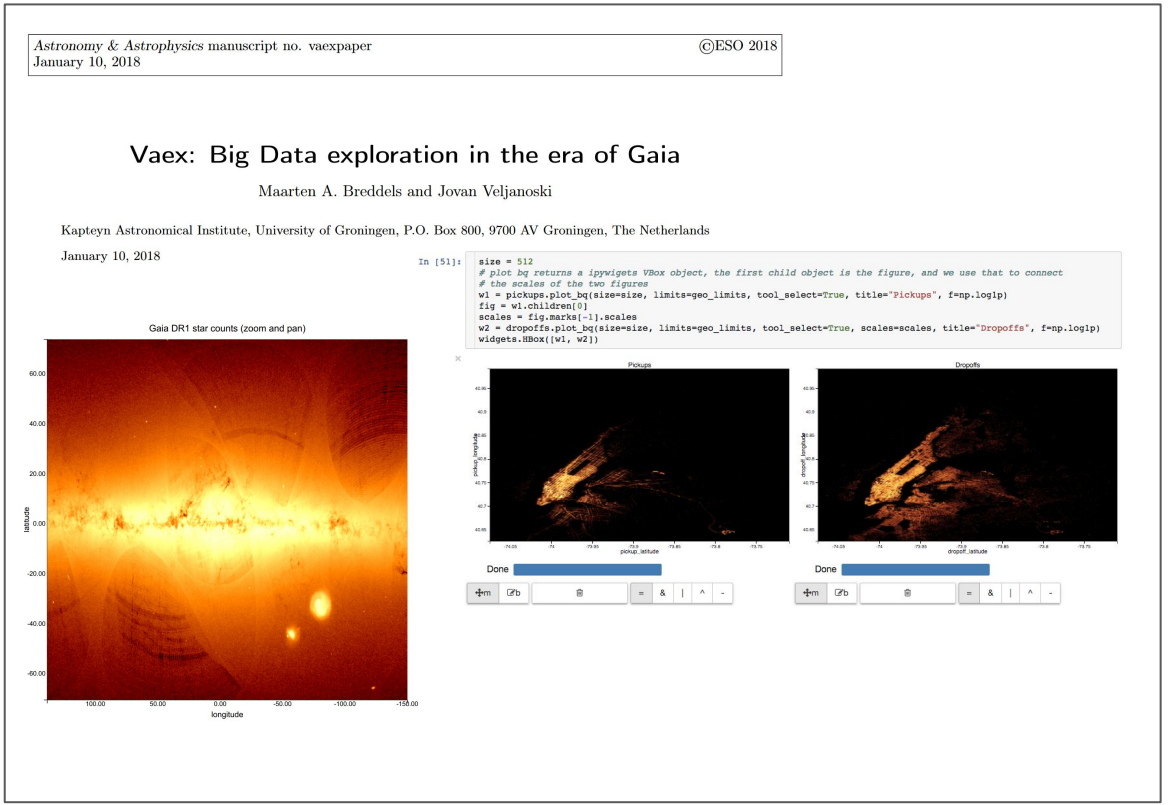
Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



Jupyter in Astronomy

- Becoming standard practice to publish notebooks along with papers, including LIGO results (and many others)
- One of the most common tools used by Astronomers to do analysis
- ... and education
- Multi user interface adopted by many projects (DES, LSST, NASA, STScI, NOAO, SciServer, etc)
- Tools and extensions developed by/for astronomers



Astronomy & Astrophysics manuscript no. vaexpaper
January 10, 2018

©ESO 2018

Vaex: Big Data exploration in the era of Gaia

Maarten A. Breddels and Jovan Veljanoski

Kapteyn Astronomical Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

January 10, 2018

```
In [51]: size = 512
         # plot bq returns a ipywigets VBox object, the first child object is the figure, and we use that to connect
         # the scales of the two figures
         w1 = pickups.plot_bq(size=size, limits=geo_limits, tool_select=True, title="Pickups", f=np.log1p)
         fig = w1.children[0]
         scales = fig.marks[-1].scales
         w2 = dropoffs.plot_bq(size=size, limits=geo_limits, tool_select=True, scales=scales, title="Dropoffs", f=np.log1p)
         widgets.HBox([w1, w2])
```

Gaia DR1 star counts (zoom and pan)

latitude

longitude

Pickups

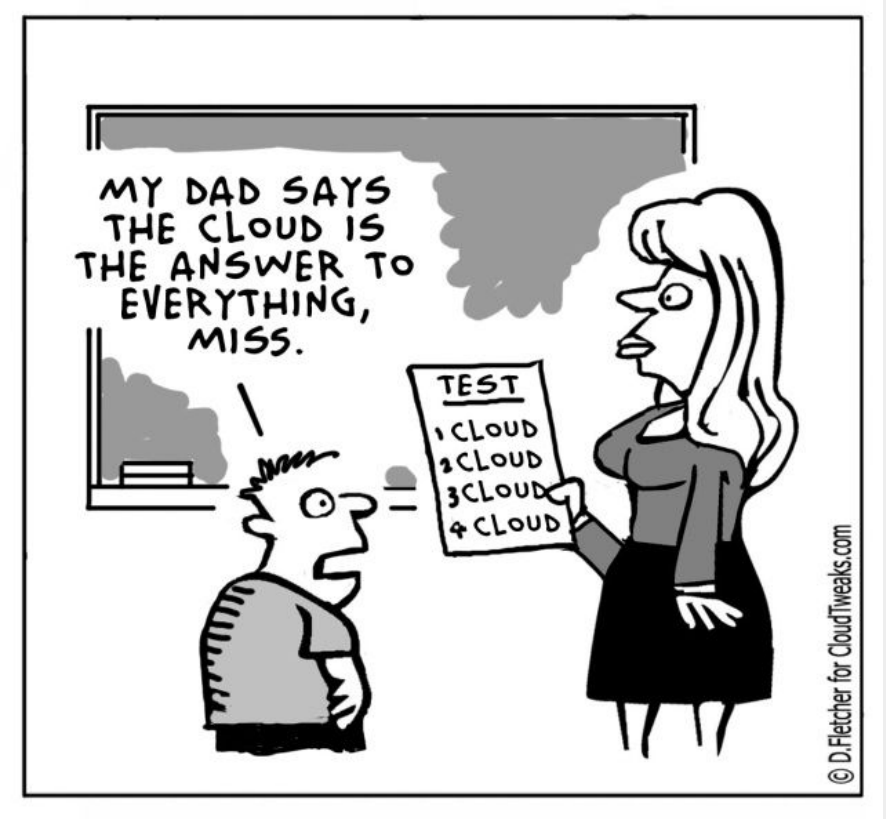
Dropoffs

Done

Done

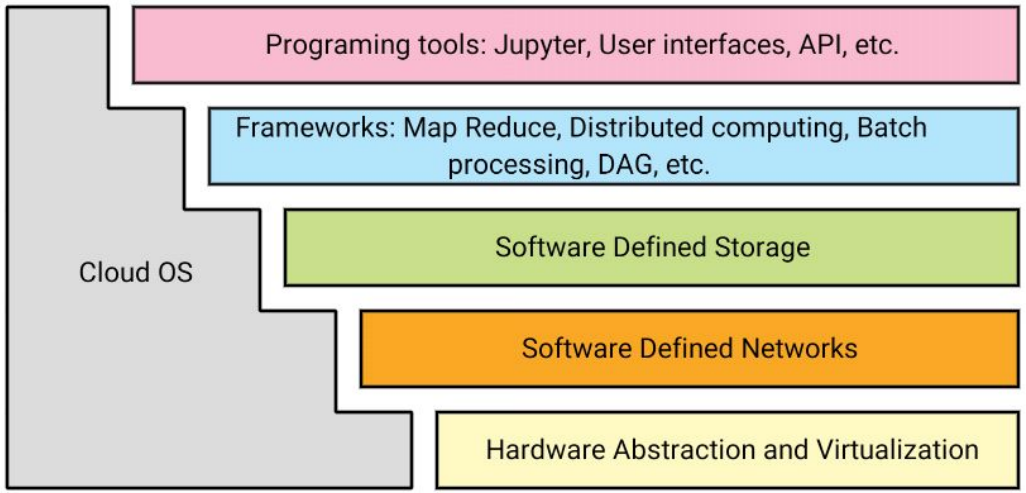
Scientific Cloud Computing

Cloud is about how you do computing, not where you do computing.



Why we should be doing science on the cloud

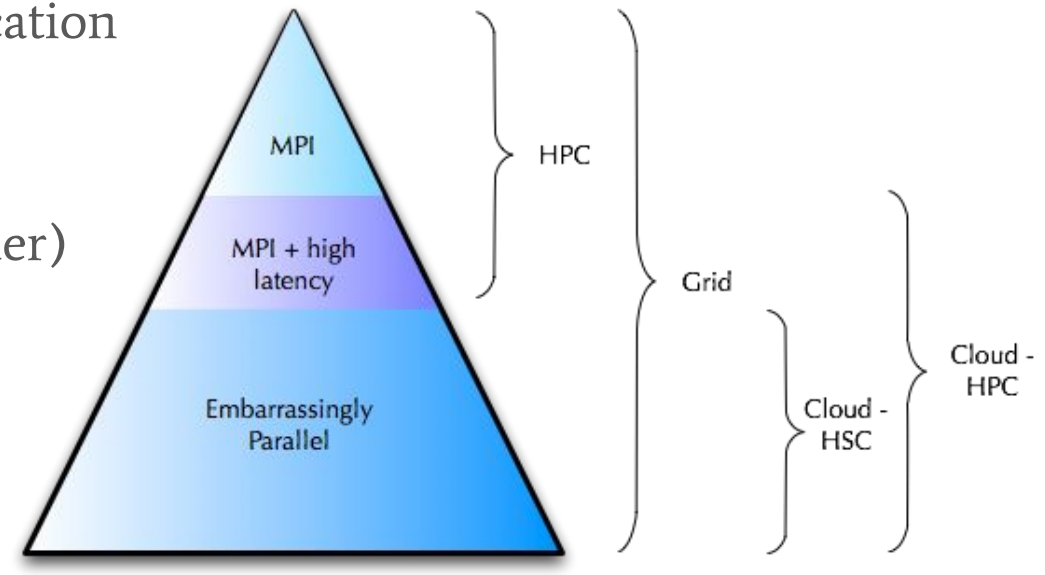
- Remote and dynamic data (!= Big data)
- Big data ⇒ Data Gravity, Data Lake, etc.
- Remote software/server
- Easy to deploy*
- Asynchronous
- Web applications / Shareable
- Serverless applications
- Federation of Services
- Tablets/ChromeOS
- more...



*arguable

Why we shouldn't be doing science on the cloud

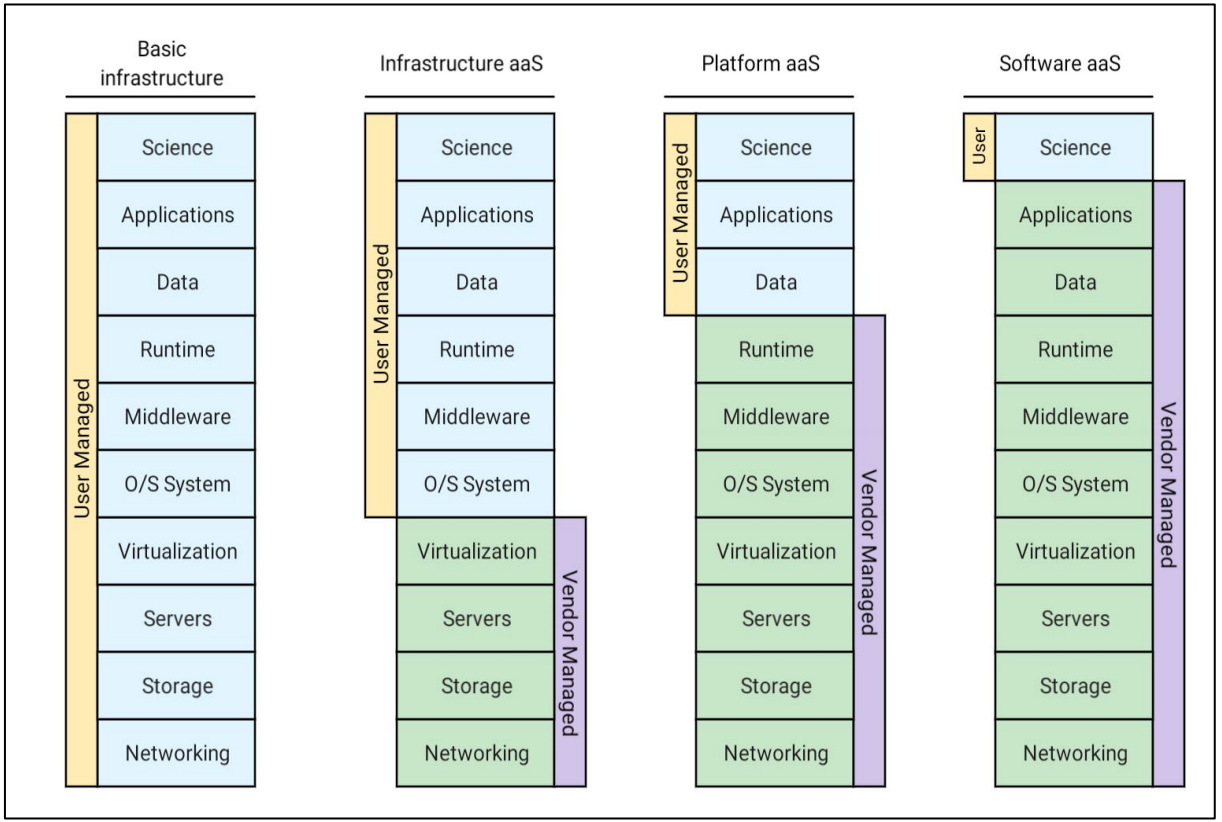
- Because there is no a real reason for it^
- HPC is not there yet, large latencies and bad bisection bandwidth
  ... but HPC is adopting cloud technologies
- Full control on data and application
- Security concerns
- Faster development*
- Billing (if a commercial provider)
- more ...



^arguable
*arguable (CI, CD)

What kind of science/projects? → Which model

- HTC vs HPC vs HSC
- Interactive
- Small projects
- Visualizations
- Short term projects*



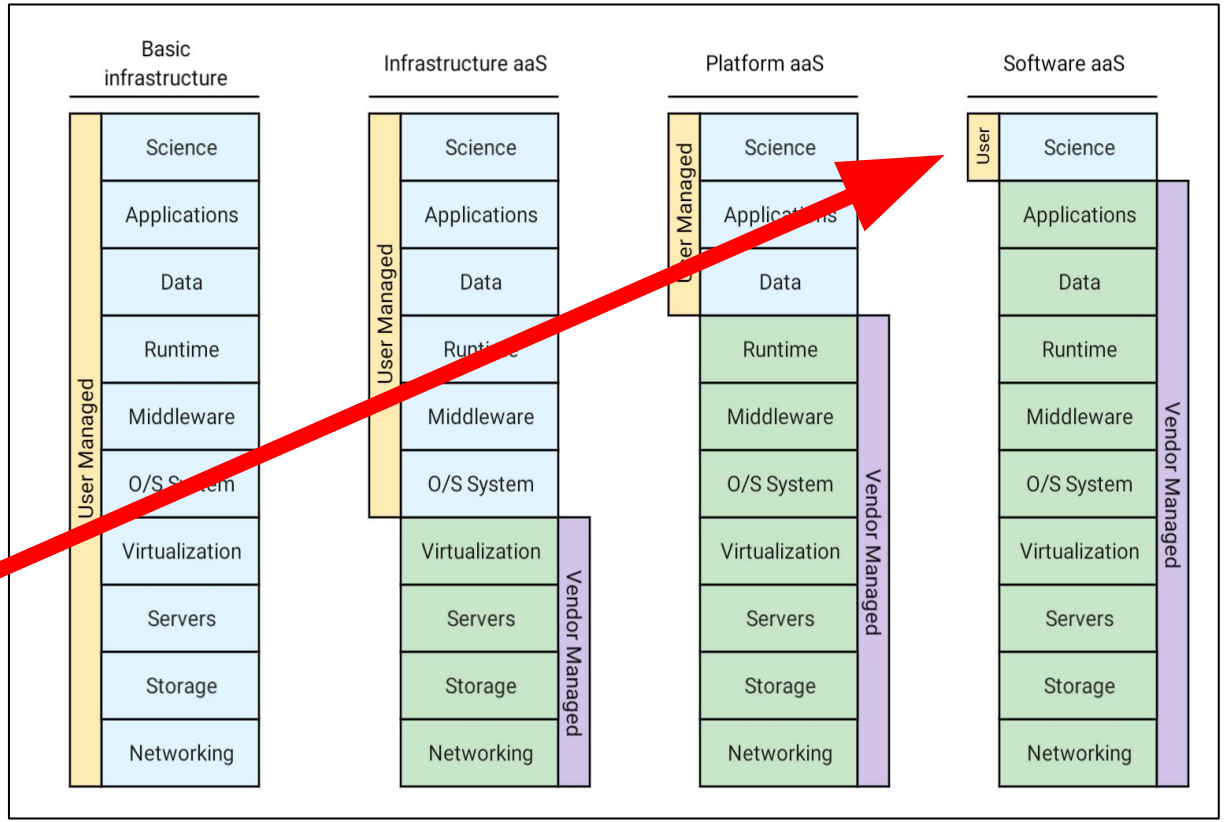
*arguable

What kind of science/projects? → Which model

- HTC vs HPC vs HSC
- Interactive
- Small projects
- Visualizations
- Short term projects*

Will we get to have Science as a Service (SCIaaS?)

*arguable



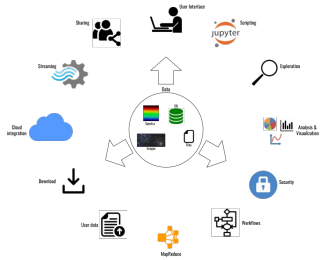
Which Clouds?

Amazon Web Services (AWS) – 40%
Microsoft Azure – about 50% of AWS
Google Cloud – 3rd place
IBM Bluemix – growing VERY fast

Salesforce, DigitalOcean, Rackspace, 1&1, UpCloud, CityCloud, CloudSigma, CloudWatt, Aruba, CloudFerro, Orange, OVH, T-Systems

Cloud for Research: Aristotle, Bionimbus, Jetstream, Chameleon, RedCloud

Why not both?



Scale on demand

User AWS account

Offered services

Processing

Local

Cloud

OSG

HPC

VMs

Remote storage

User Applications

Applications, services and user interfaces

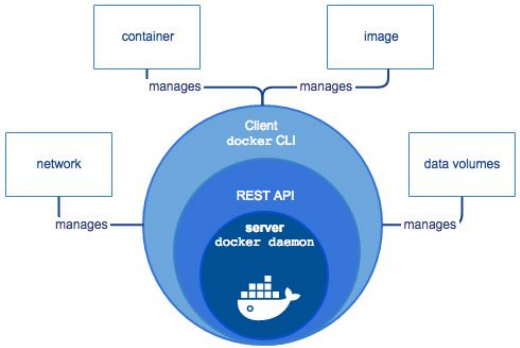
NFS, GridFTP, Globus, etc

Edge VMs, Services

Storage and Storage Services

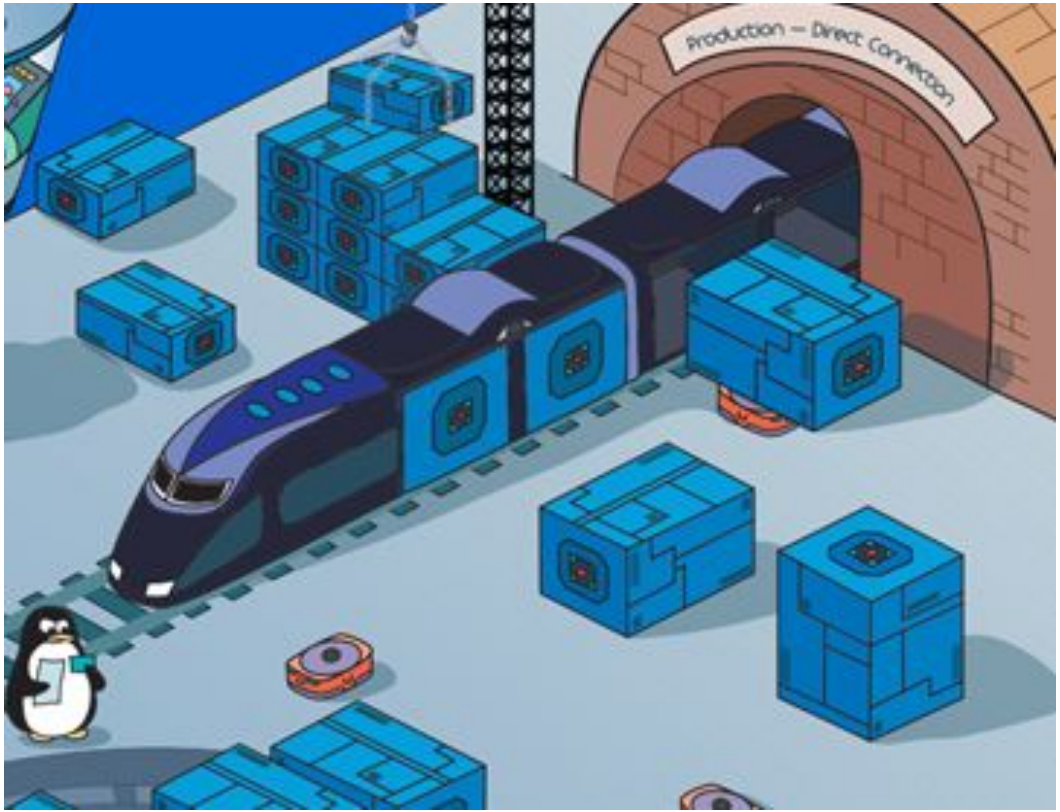
Containerization to the rescue

- It's been around for over 10 years, but popular since 2014 thanks to Docker
- Many other alternatives (rkt, kata, shifter, singularity, etc...)
- Lightweight, stand-alone, executable package of a piece of software that includes everything to run it
- Not just applications
- Software designed storage
- Software designed network



Container organization and orchestration

- We can create a container with an application inside, now what?
- Need to consider:
  - Resource needs
  - Fault tolerant
  - Load balancing
  - Storage management
  - Lifecycle
  - Service Discovery
  - Scalability



The Kubernetes Factor

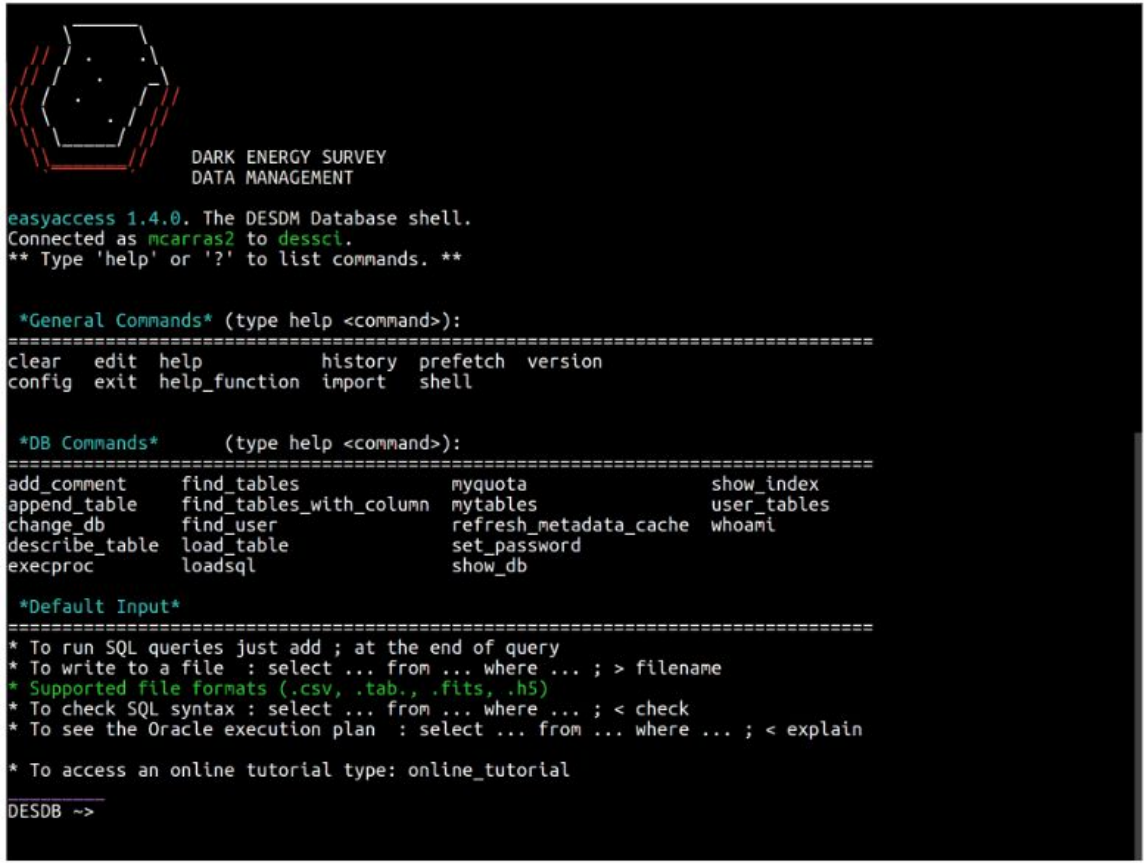
- It solves all previous issues and more (not the only one but most popular)
- Open source container management and orchestration platform
- Developed by Google, made open sourced
- One of top 5 most commented open source repositories and #2 in number of pull request
- Standard within all cloud platforms
- Flexible and extensible, customize schedulers
- Is changing the cloud computing paradigm

Applications

The Dark Energy Survey

- 4 meters telescope, 520 Mpx camera
- 5 year survey, ⅛ of the sky, Telescope in Chile, data @ NCSA, about to start 6th season
- Main Goal: To constrain the models of the Universe regarding Dark Energy and Dark Matter.
- Many other Science Cases! (New dwarf planet, New galaxy satellites, Supernovae, etc)
- 1 - 3 TB of data per night, 1 PB of data
- Processing done at FermiGrid, Campus Cluster and Blue Waters
- Thousands of images and billions of rows, ~500 millions objects
- 1st Public Data Release in January 2018
- NCSA provide means to access and interact with data → Containers

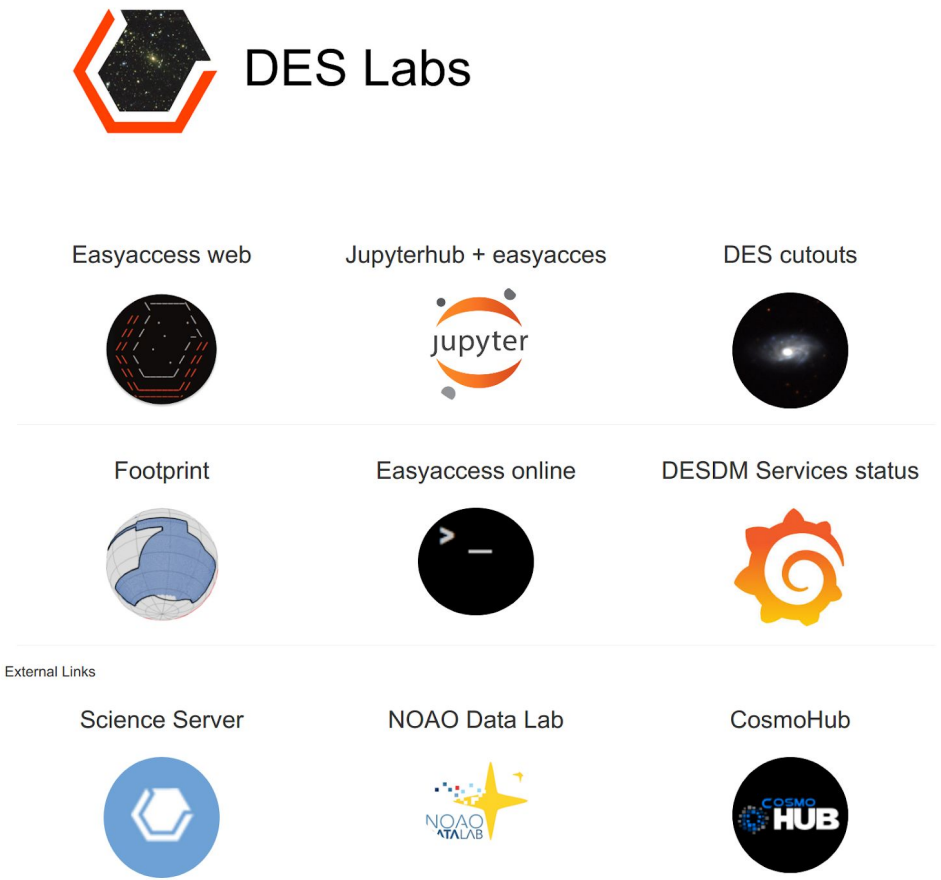
easyaccess: DES command line tool



- DES DB in Oracle
- Specifically designed for DES (internal and public)
- Enhanced SQL command line interpreter in Python
- Astronomer friendly
- Python API, web interface
- There are many other CLI and GUI clients.
- Needed a simple tool, easy to use and install
- Autocompletion
- Load/Save to hdf5, fits, csv

DES Labs: Collection of containerized tools for DES

- Launched March 2015
- Used by the Collaboration
- Running using Kubernetes at NCSA cloud
- Currently being migrated to match DR1 Infrastructure



NCSA DESaccess: Services

DARK ENERGY SURVEY desaccess

mcarras2
mcarras2@ncsa.illinois.edu

- Home
- DB access
- DES Table Schema
- Example Queries
- Cutouts Service
- DES JupyterLab
- Finding Chart
- DES Footprint
- Data Analysis
- My Jobs
- Help

Welcome to dessci, Matias!

DB ACCESS

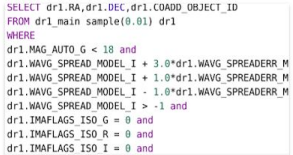
Oracle SQL web- client

More...

DES TABLE SCHEMA

Browse all tables

More...



EXAMPLE QUERIES

See some example queries as a start

More...

CUTOUTS SERVICE

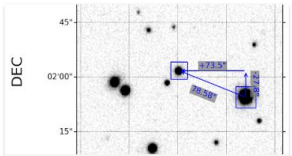
Generate cutouts for positions or ids

More...

DES JupyterLabs

(Beta) Jupyter Labs

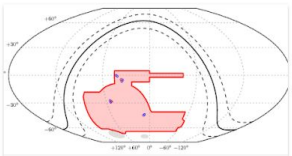
More...



FINDING CHART

Find your object

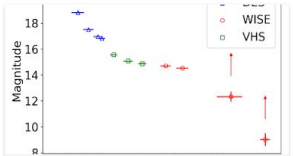
More...



DES FOOTPRINT

Interactive globe

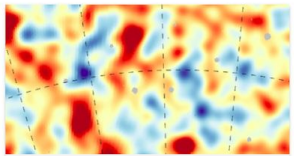
More...



DATA ANALYSIS

SEDs and color-color diagrams

More...



MY JOBS

List of submitted jobs

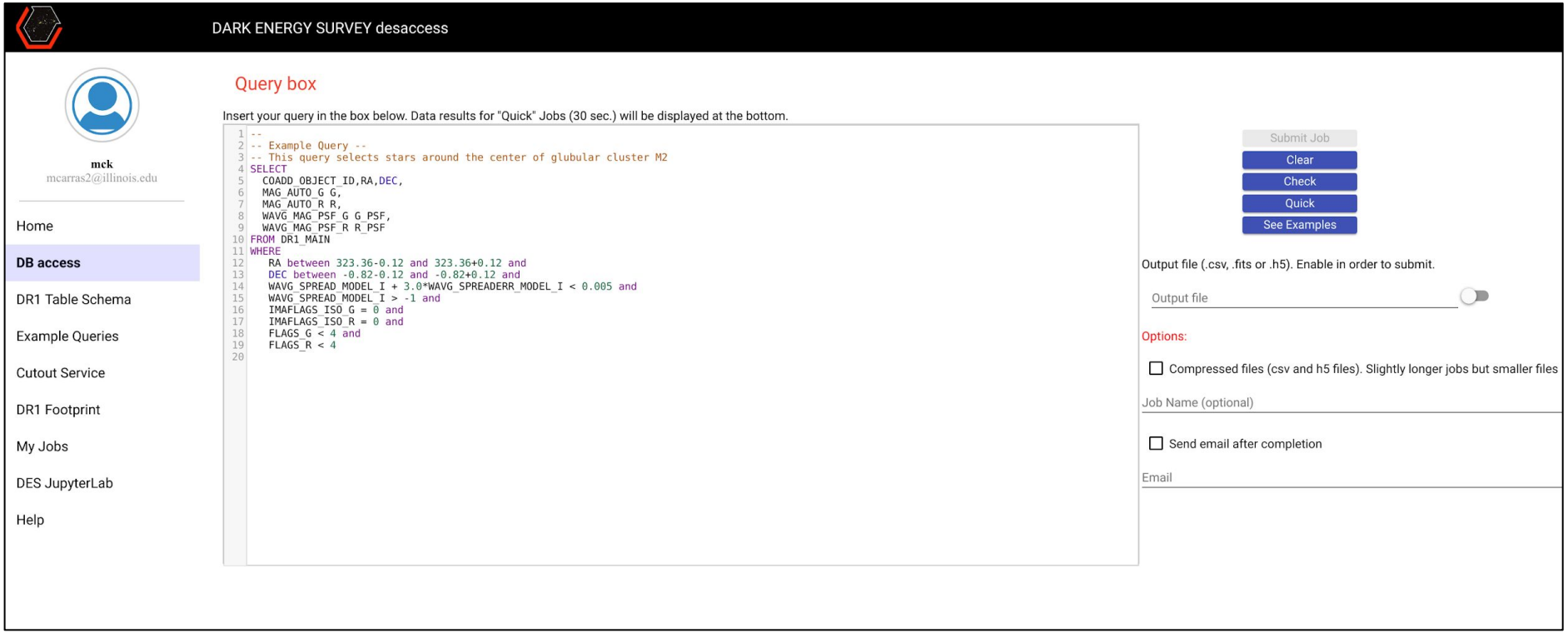
More...

HELP

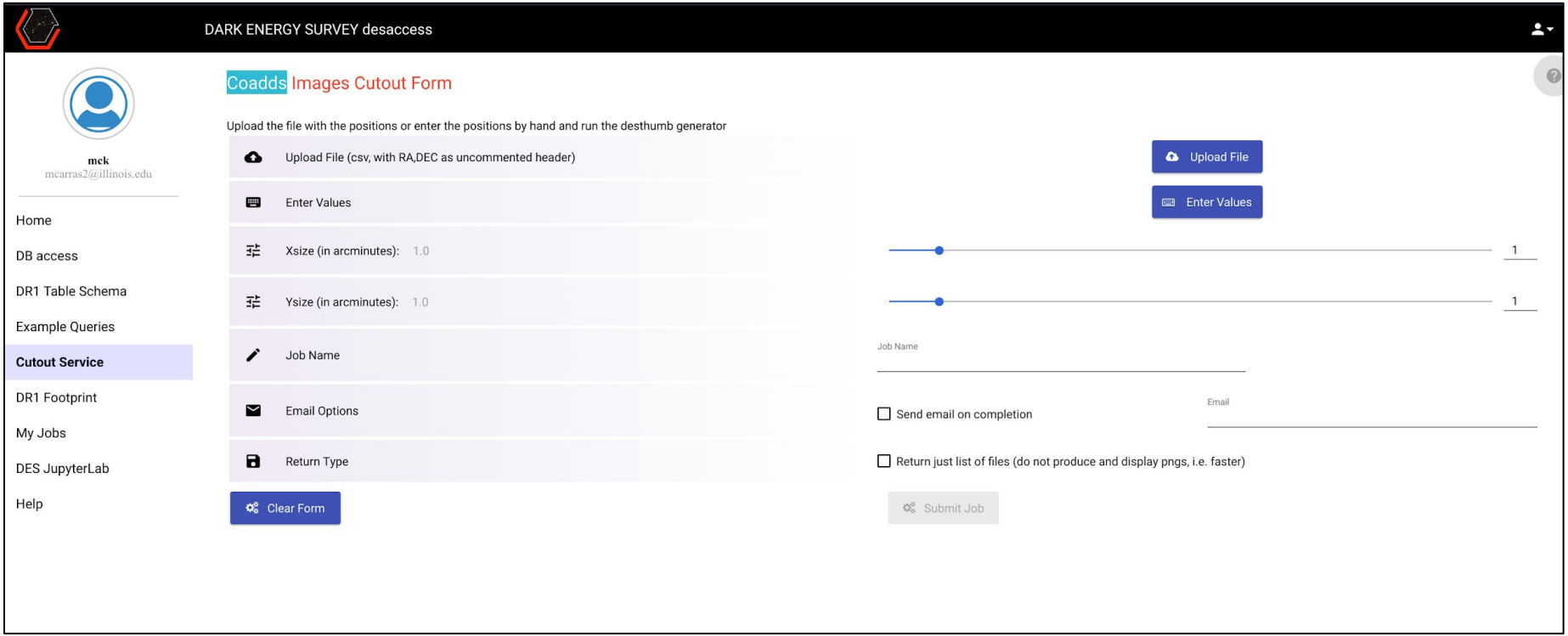
Help form

More...

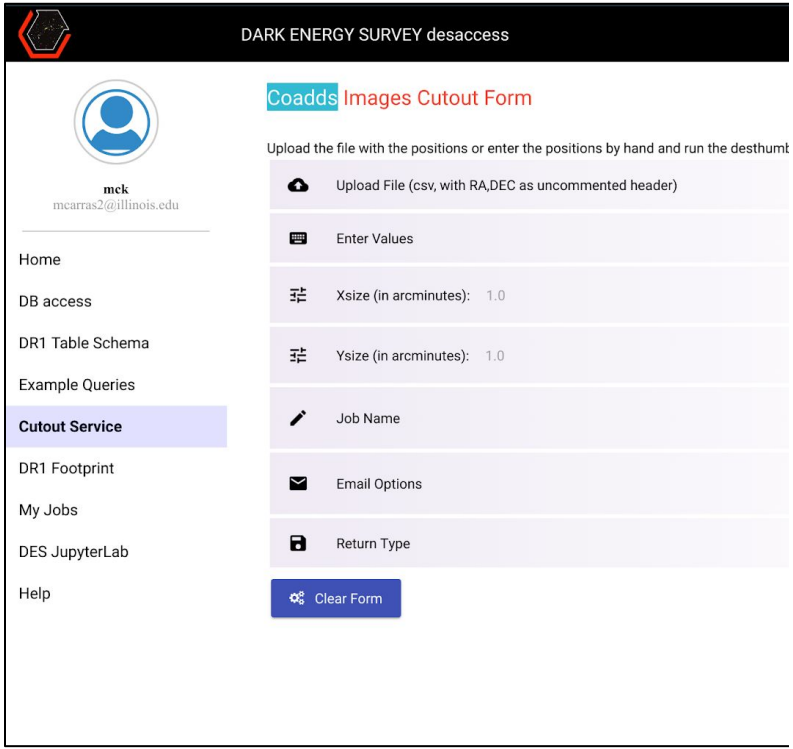
NCSA DESaccess: DB access



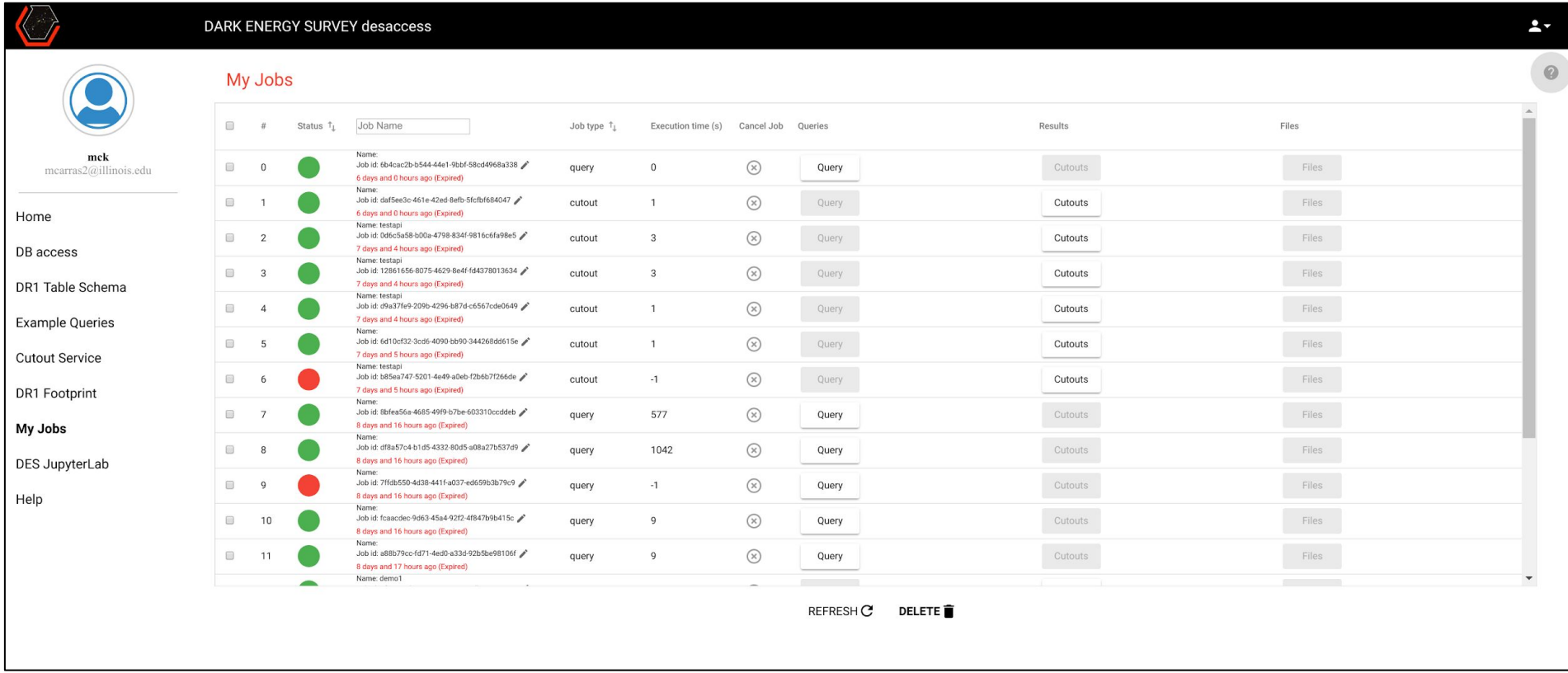
NCSA DESaccess: Cutouts Service



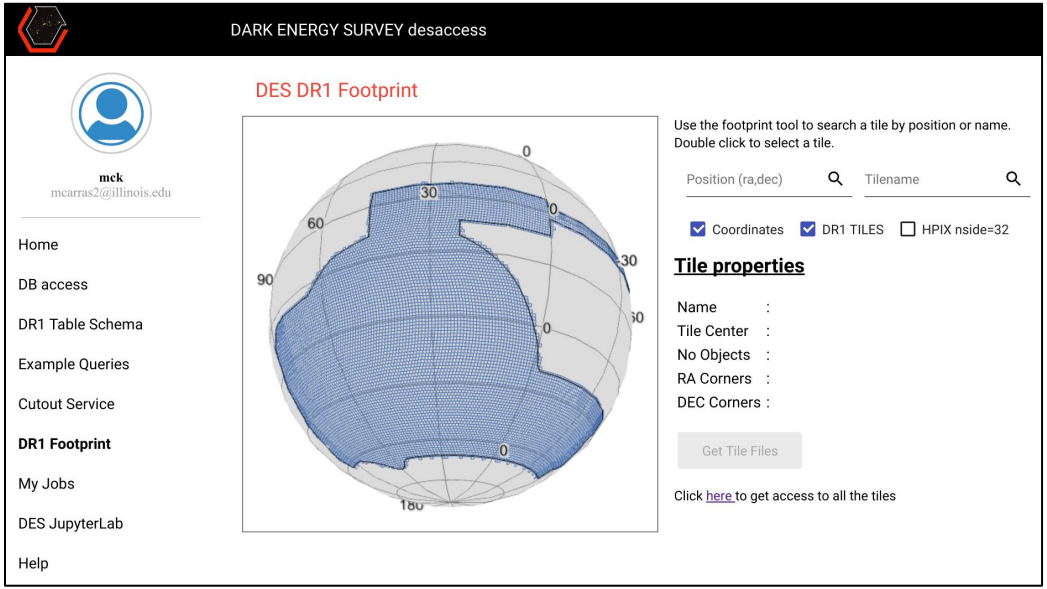
NCSA DESaccess: Cutouts Service

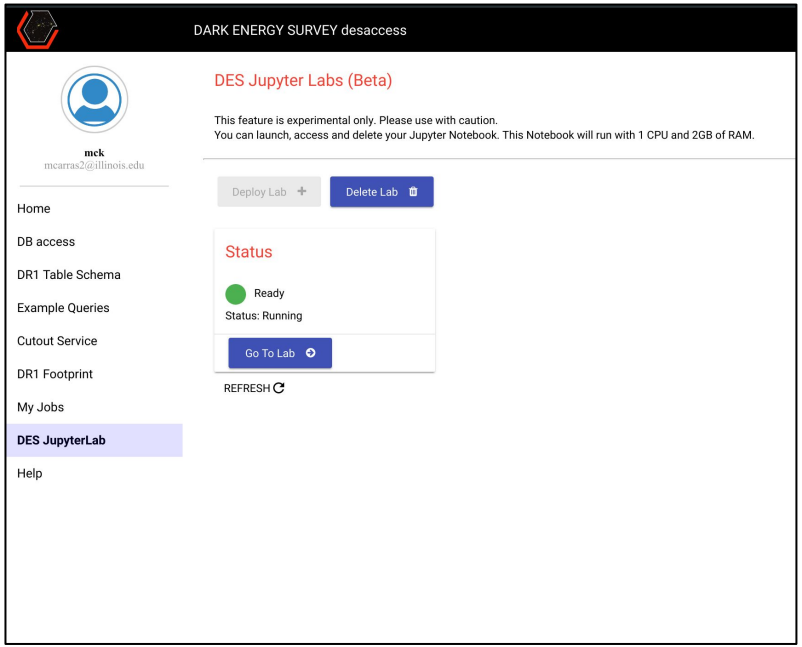


NCSA DESaccess: Asynchronous Jobs

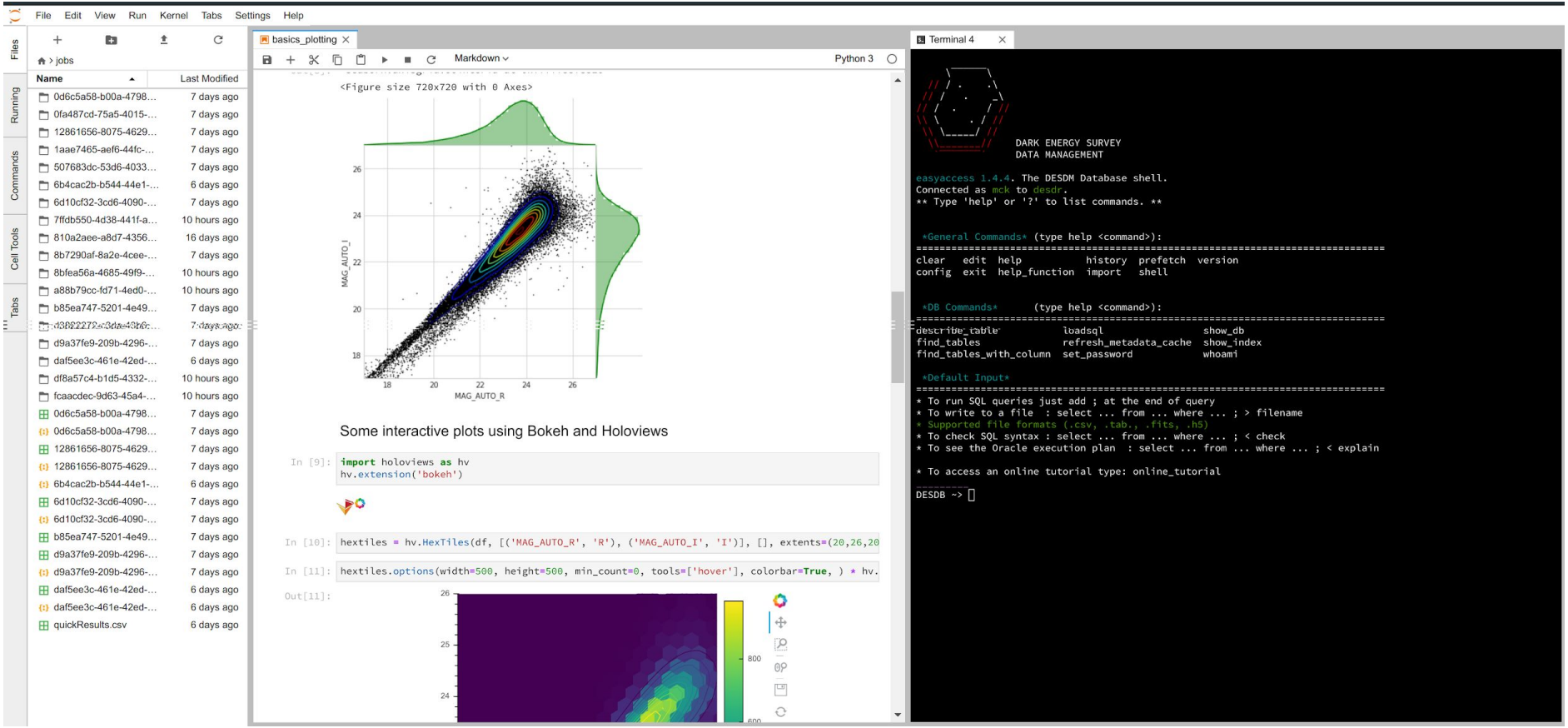


NCSA DESaccess: Footprint and Jupyter Labs

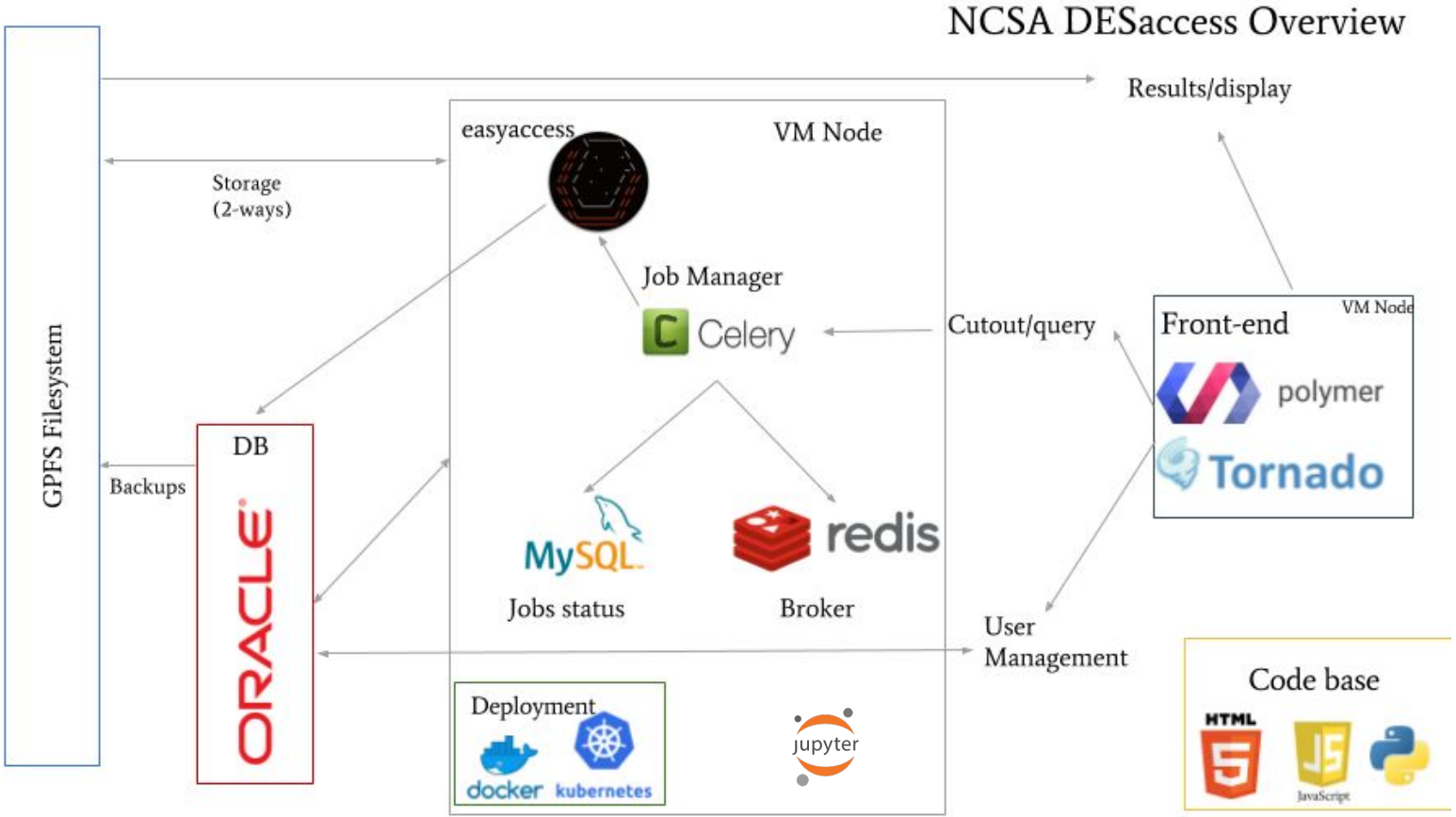




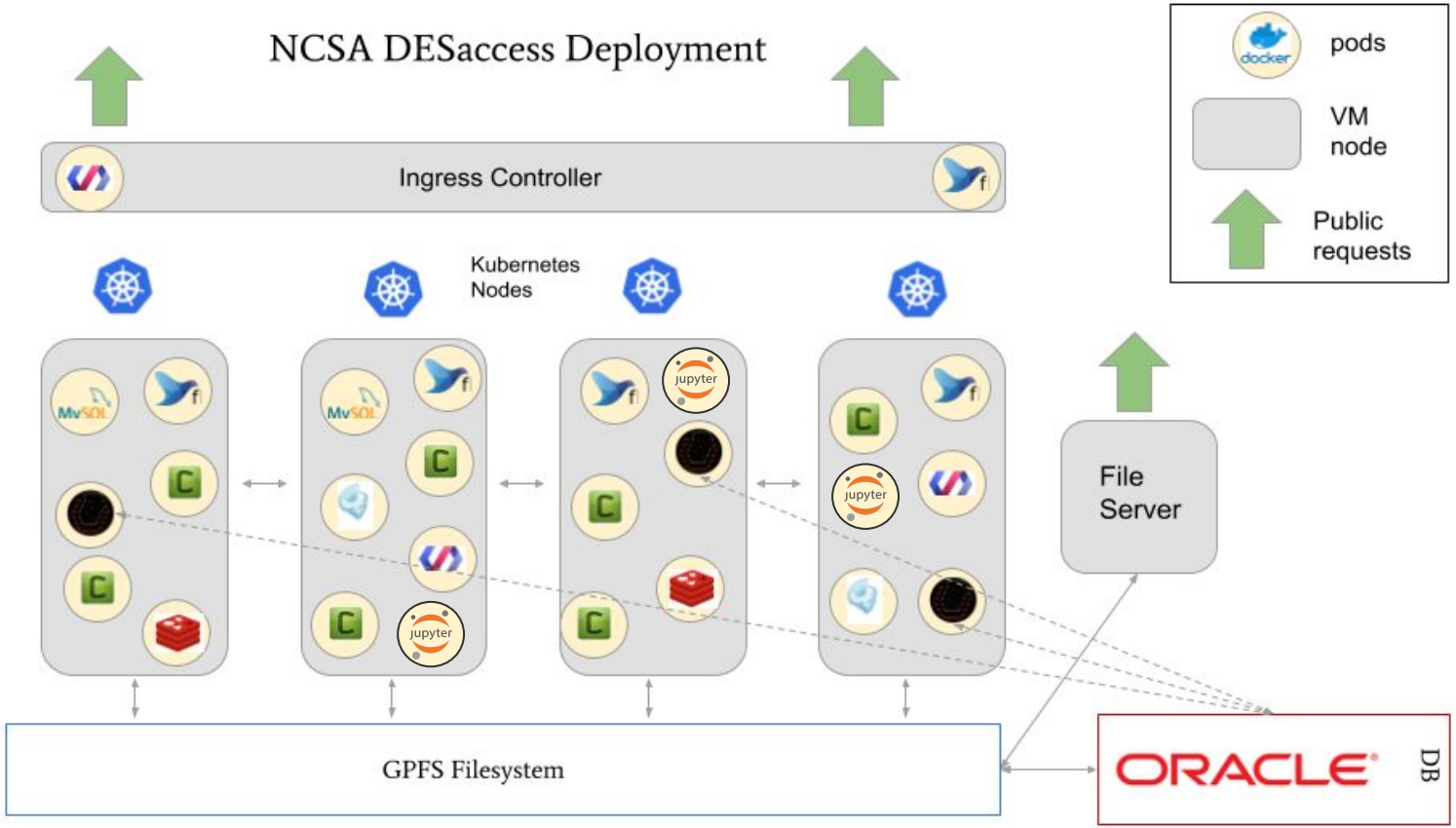
NCSA DESaccess: Labs with access to Jobs and easyaccess



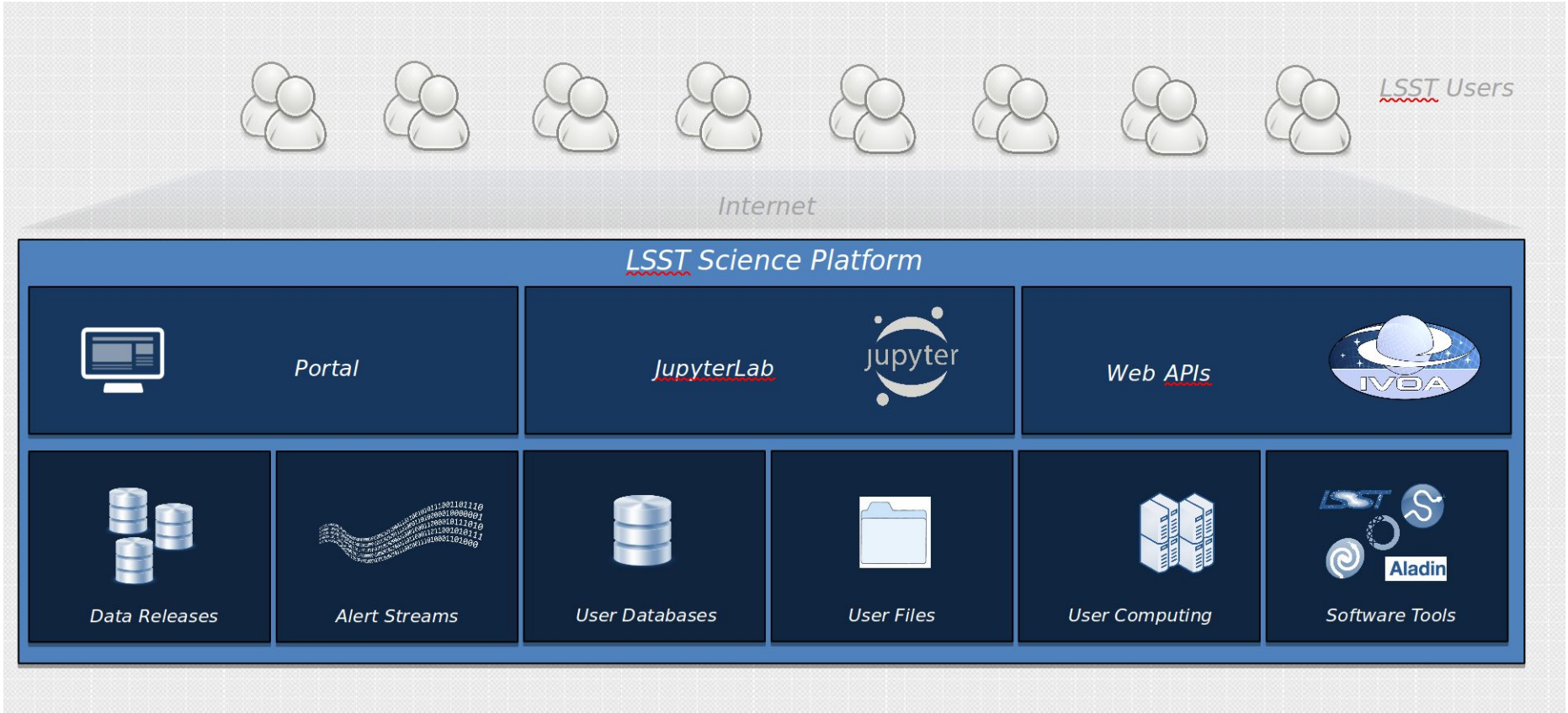
NCSA DESaccess: Technology Overview



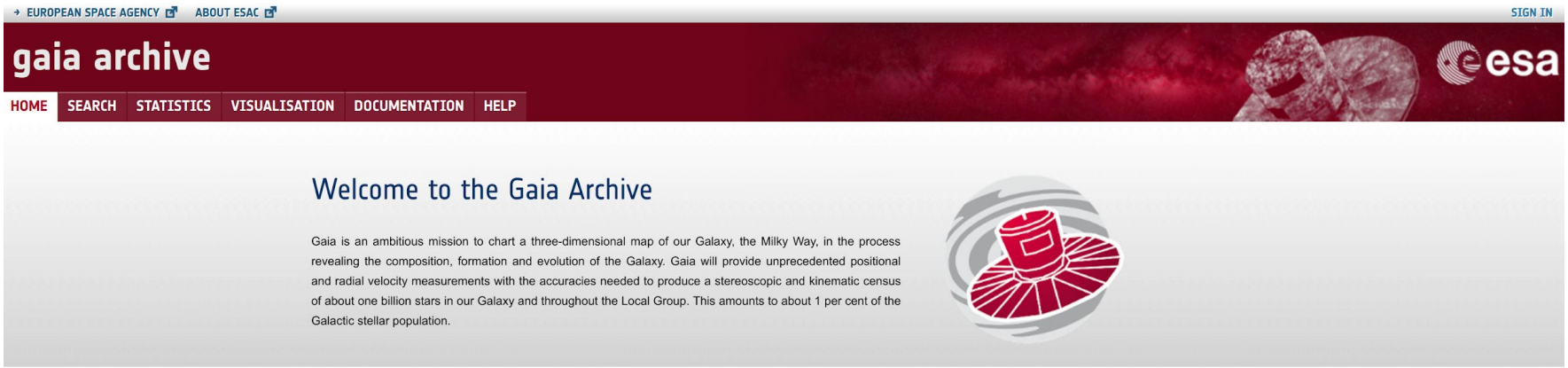
NCSA DESacces: Deployment



LSST Science Platform



GAIA Archive



SciServer


SciServer
Collaborative data-driven science
A new vision for science
A collaborative research environment for large-scale data-driven science
SciServer Betelgeuse v2.0.3
Login to SciServer
About
Bringing the Analysis to the Data
Tools
A modular system of independent components.
Hosted Datasets
SciServer makes a number of datasets available online in the…
Science
Full datasets
Common formats
Common sets of interfaces
Education
Building relationships with universities, institutes, and government organizations.
Help
Workshop supporting material, online documentation, provide feedback, report bugs.
SciServer is administered by
idies
JOHNS HOPKINS UNIVERSITY
SciServer is funded by National Science Foundation award ACI-1261715
NSF

SClaaS Example: Anomaly detection service

Goal: Build a resilient scalable anomaly detection service.

Motivation: Astronomical data (both literal and figurative)

Algorithm: Extended Isolation Forest

Infrastructure: Kubernetes cluster

MapReduce package: Spark

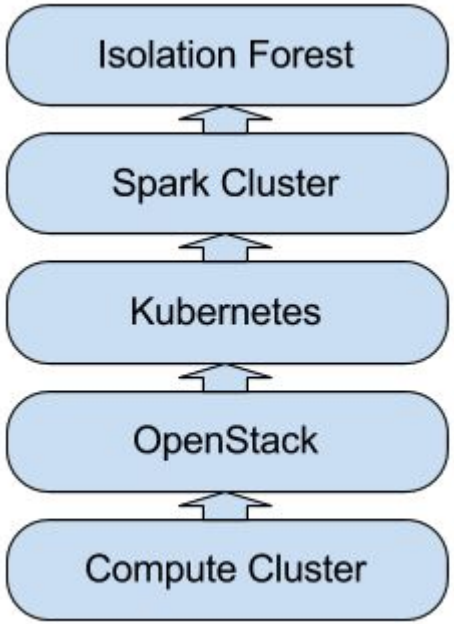
Technology Stack For Anomaly Service

Batch and online anomaly detection for scientific applications in a Kubernetes environment

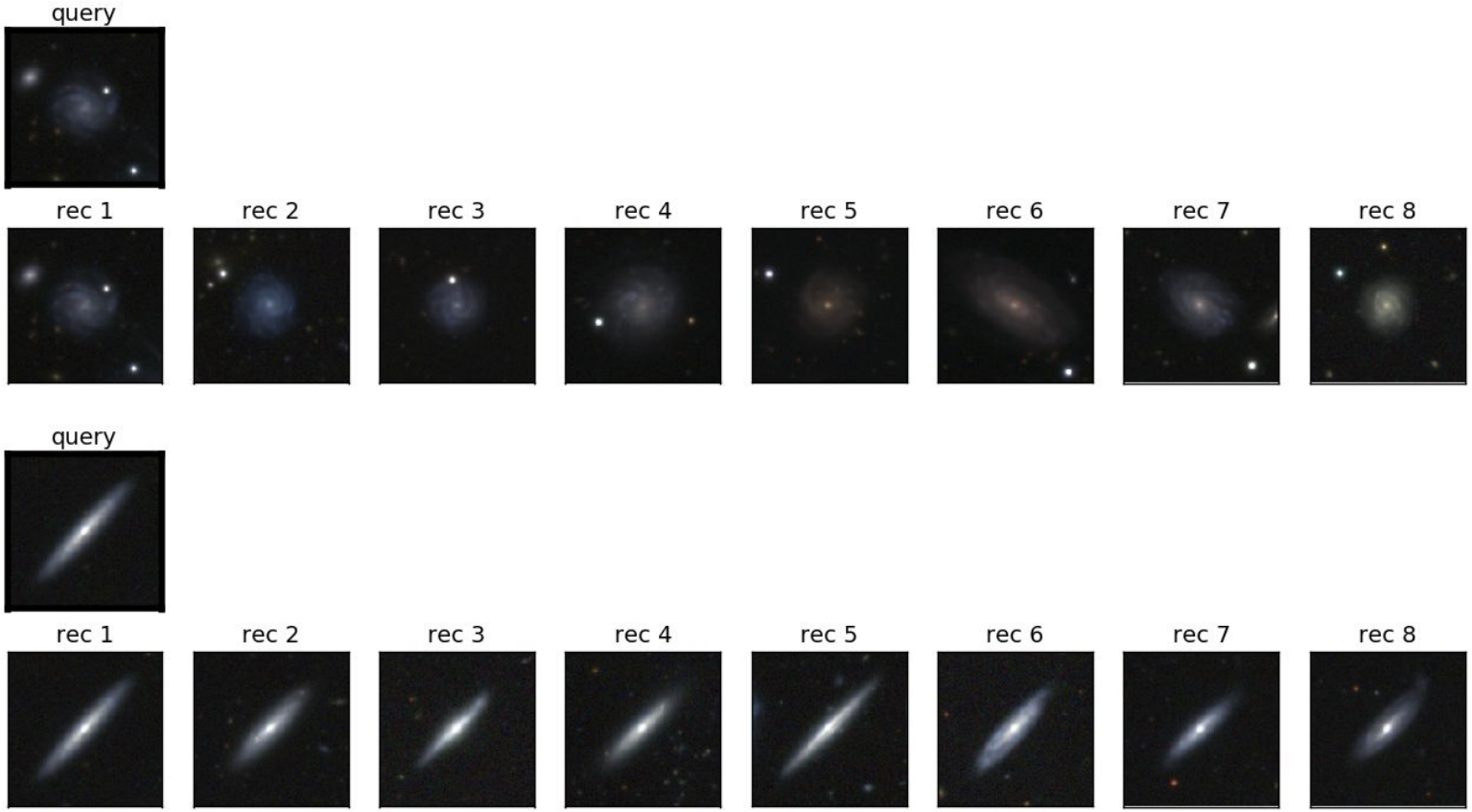
Sahand Hariri*
University of Illinois at Urbana-Champaign
sahandha@gmail.com

Matias Carrasco Kind†
National Center for Supercomputing Applications
mcarras2@illinois.edu

- Use Extended Isolation Forest as core algorithm
- Use Spark to parallelize trees and scoring
- Use Redis as a broker communicator
- To easily deploy in any environment, use Docker
- For orchestration of Docker containers, use Kubernetes
- Kubernetes cluster built on top of OpenStack, but it can be deployed also in AWS, GKE, etc.



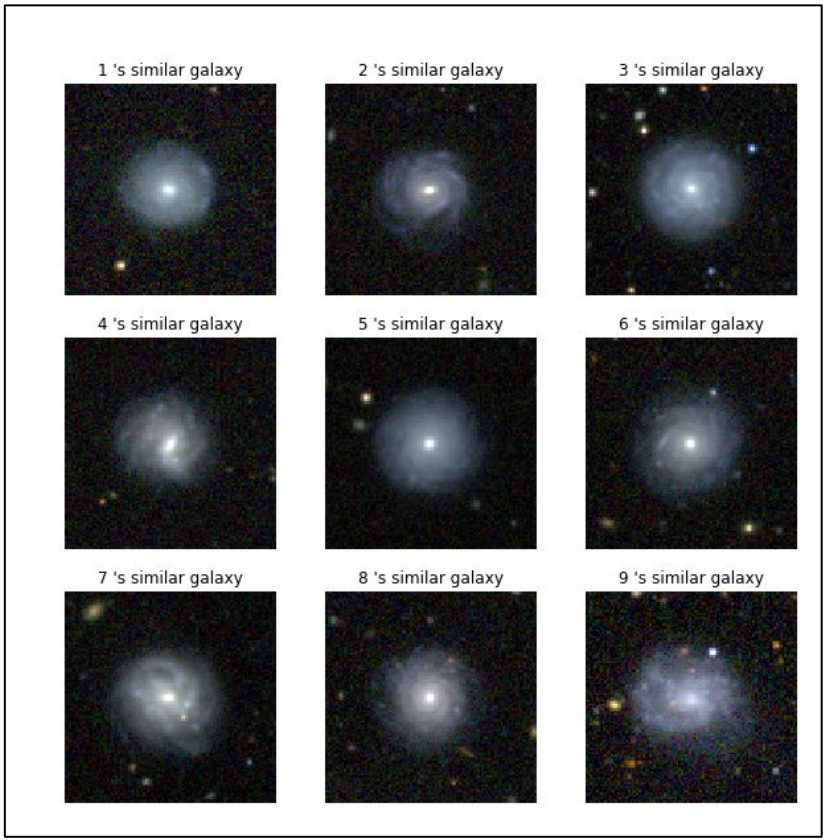
SClaaS Example: Galaxy selection and similarity search

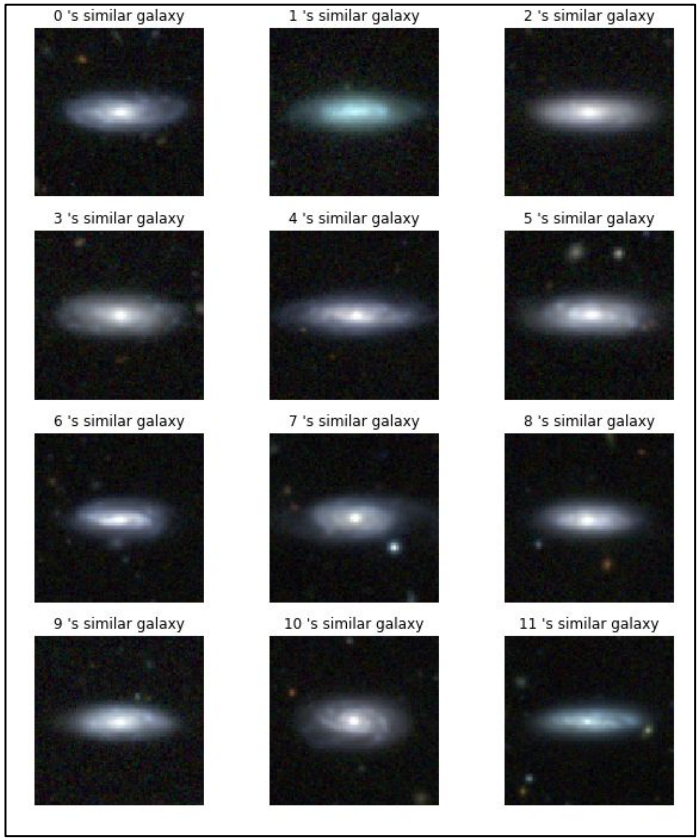


Compress images from 200x200 to 50 or less, for fast search

Bowen, Carrasco-Kind, et al. in prep.

SClaaS Example: Galaxy selection and similarity search





Bowen, Carrasco-Kind, et al. in prep.

Final Remarks

- It's all about the user
- Jupyter as Scientific tool
- Science on the cloud is happening in many scientific fields including Astronomy
- Containerized solutions to ease management of the applications
- HPC is adopting cloud technologies to leverage the benefits of both worlds
- Kubernetes provide means to have 'the cloud' outside the commercial world
- Production services for large datasets
- YOU are not alone

... this is changing the way we do astronomy

Matias Carrasco Kind -- NCSA
mcarras2@illinois.edu
github.com/mgckind
matias-ck.com

Thank you!

Questions?

Matias Carrasco Kind -- NCSA
mcarras2@illinois.edu
github.com/mgkind
matias-ck.com