



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?

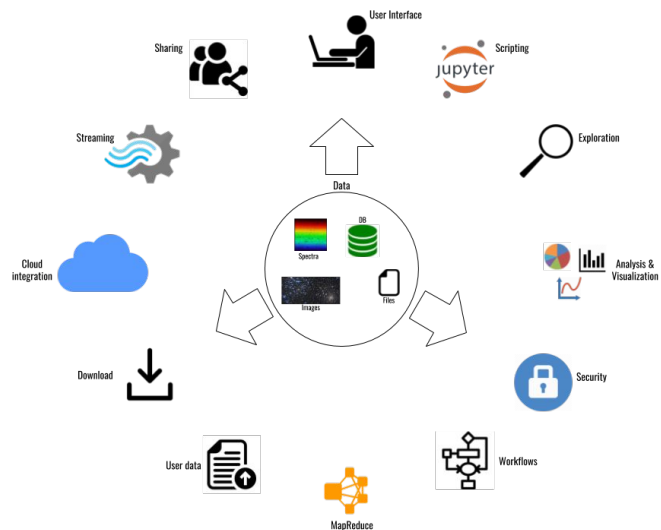


Several meanings around a central data archive, a.k.a “data lake”, repository with common components

- 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.)



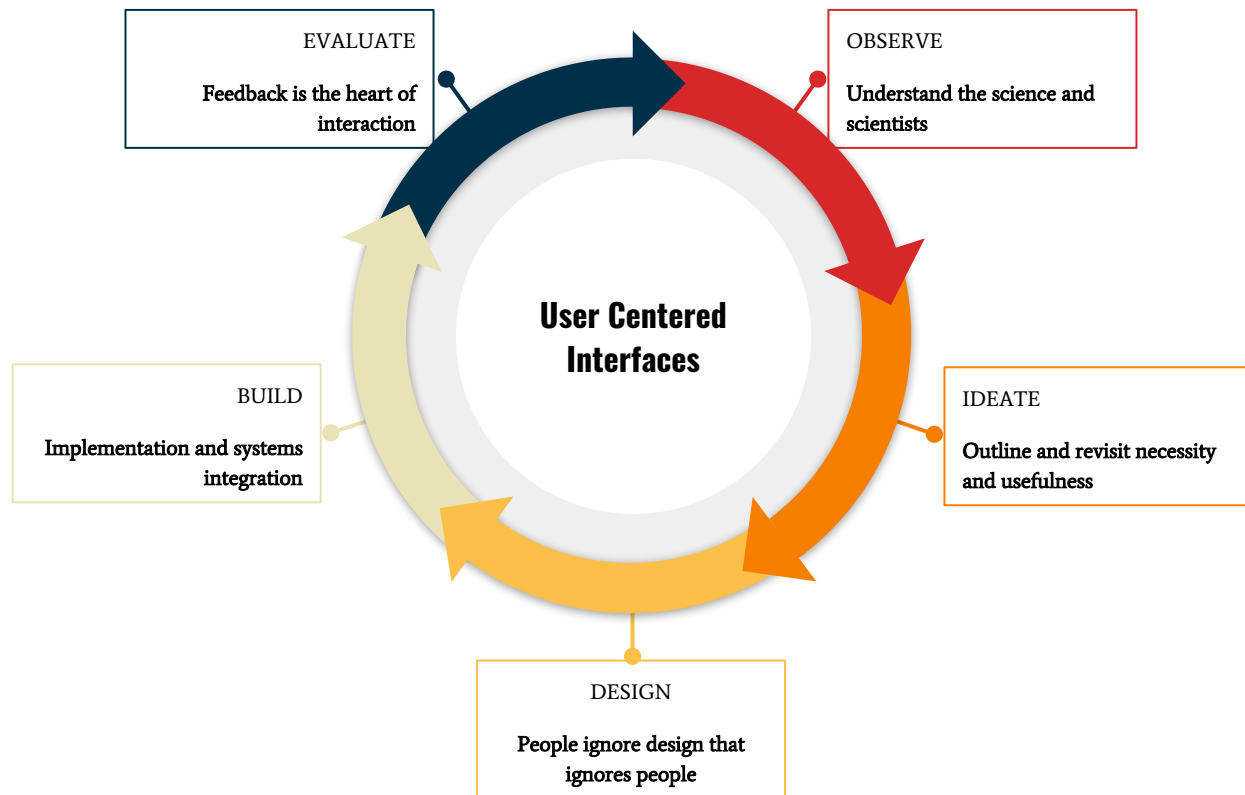
“Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines.”
(Science Gateways Institute)

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

User (Scientist) Centered Design

Data Access would not exist 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”

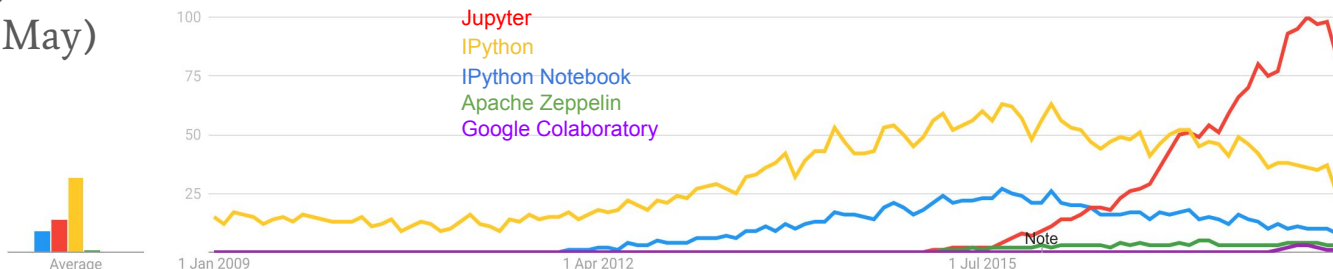
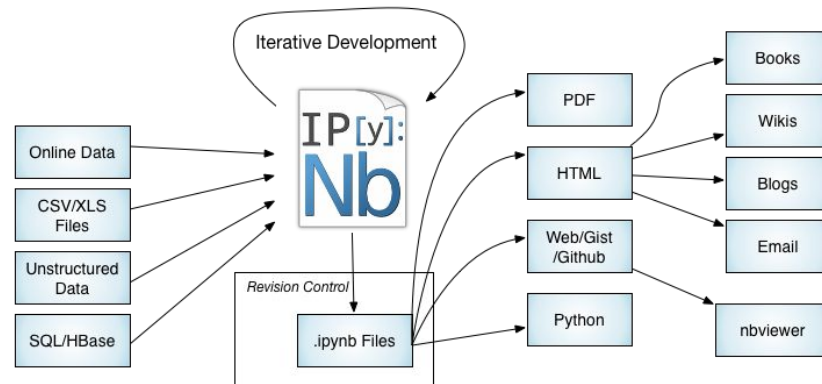


The Notebook Revolution



The Notebook Development

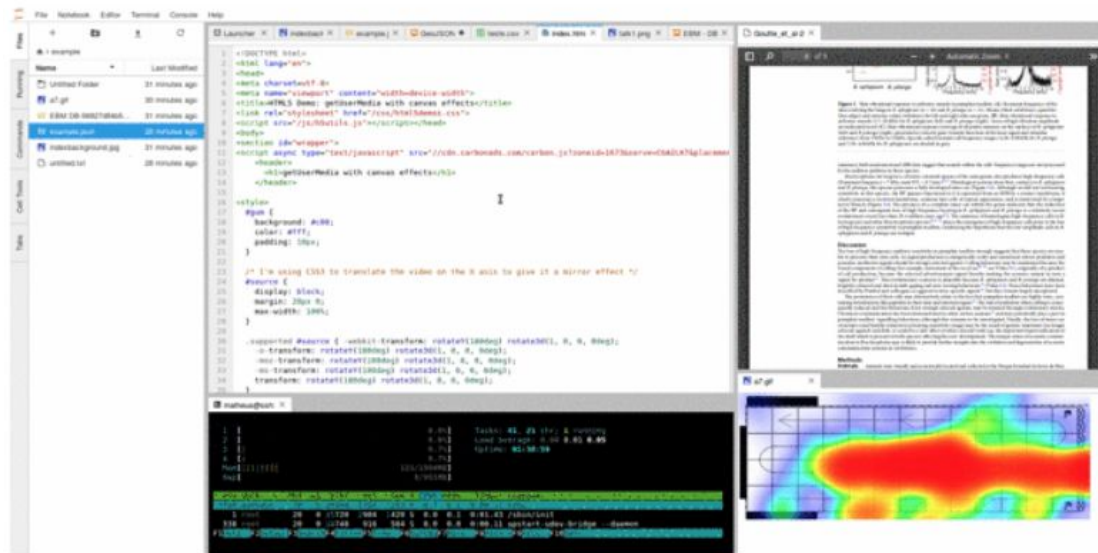
- Started from ideas like Matlab, Maple or Mathematica ~1988
- IPython has been around since 2001
- Sage Notebook released in 2005 (uses IPython)
- IPython Notebook was released in 2011
- IPython Notebook moved to Jupyter in 2014
- Apache Zeppelin created in 2015 (JVM and integrated with Apache Products)
- 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

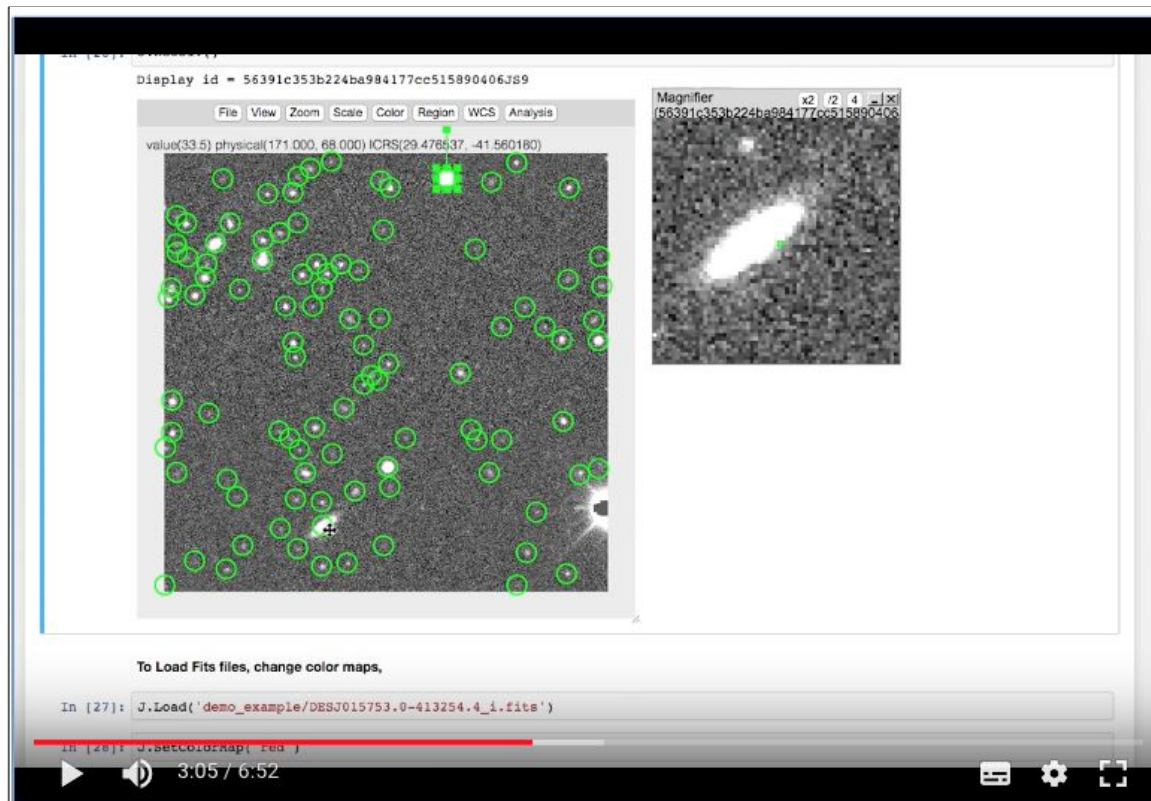


- Computational narrative
- Scripting interface
- Scientific oriented interface
- Customizable
- Collaborative
- Adopted by many projects in scientific fields
- Widgets
- Big Data Integration (Spark)
- Interactive plots
- Multiple Kernels (Python, R, Julia, Scala, etc.)



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

ELSEVIER

Astronomy and Computing
Volume 20, July 2017, Pages 128-139

Full length article
Vizic: A Jupyter-based interactive visualization tool for astronomical catalogs ☆
 W. Yu ^{a, b}, M. Carrasco Kind ^{b, c}, R.J. Brunner ^{c, b}

jupyter vizic Last Checkpoint: Last Saturday at 2:57 PM (Autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [32]: app

Color by Property

Filter Objects

Object Catalog

CLEAN	1
G	16.57518
ZS	0.00778353
PETHOP90	0.23.87925
Z	16.57518
R	16.57518
U_R	0.0073217750999999995
J	16.57518
TYPE	3
OBJID	1237962296463682600
RA	228.32743
U	16.57518
DEC	4.4774828
RADIUS	23.87925
Q_R	-0.08222153

(a) Voronoi diagram layer

(b) Minimum spanning tree layer

(c) Delaunay triangulation layer

(d) HEALPix layer. This overlay is created with `nside = 1024` and zoomed in to level 4 for a clear view of the content.

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

©ESO 2018

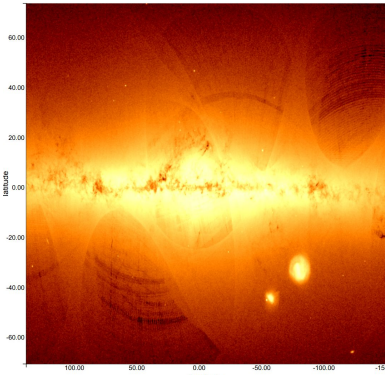
Astronomy & Astrophysics manuscript no. vaexpaper
January 10, 2018

Vaex: Big Data exploration in the era of Gaia

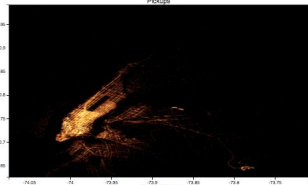
Maarten A. Breddels and Jovan Veljanoski

Kapteyn Astronomical Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands
January 10, 2018

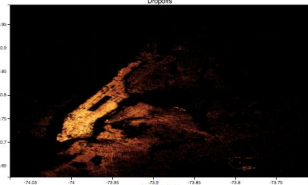
Gaia DR1 star counts (zoom and pan)



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



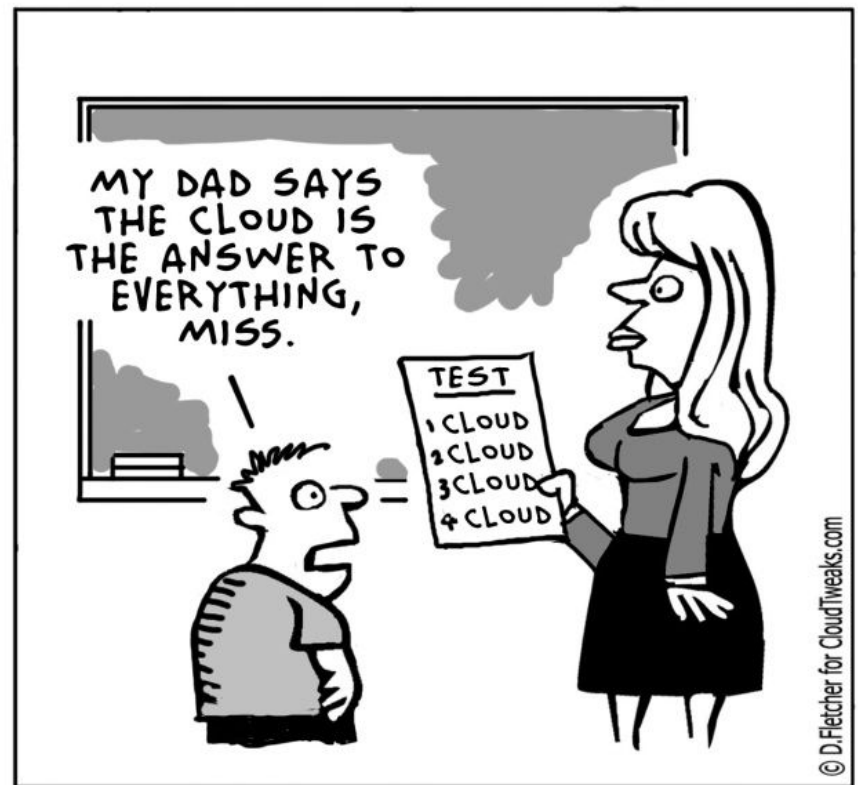
Done



Done

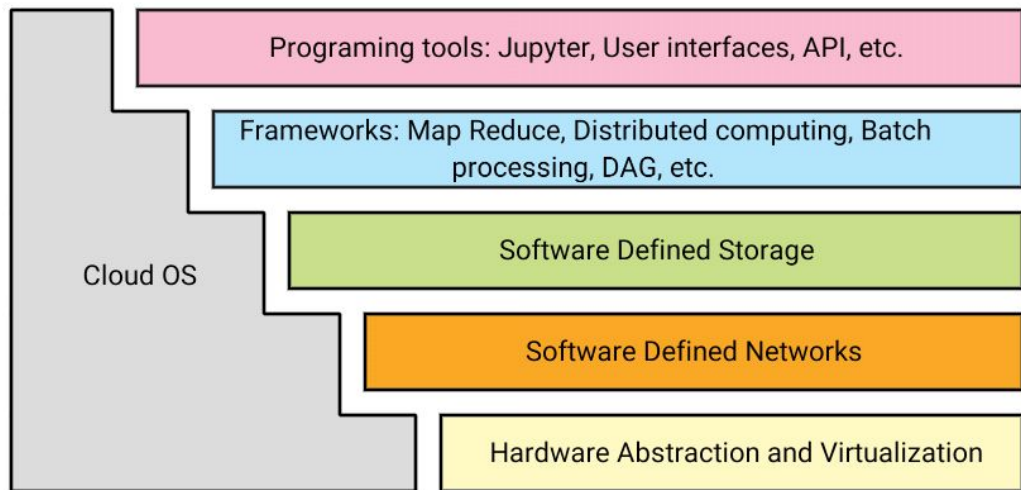
Scientific Cloud Computing

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



Why we should be doing science on the cloud

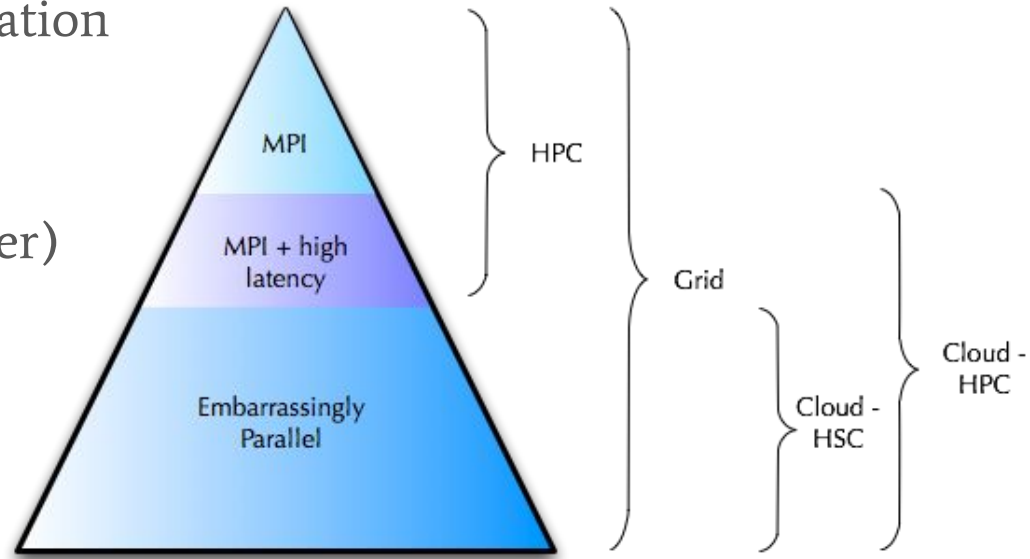
- Remote and dynamic data (!= Big data)
- Big data \Rightarrow Data Gravity, Data Lake, etc.
- Remote software/server
- Easy to deploy*
- Asynchronous
- Web applications / Shareable
- Serverless applications
- Federation of Services
- Tablets/ChromeOS
- more...



*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 ...

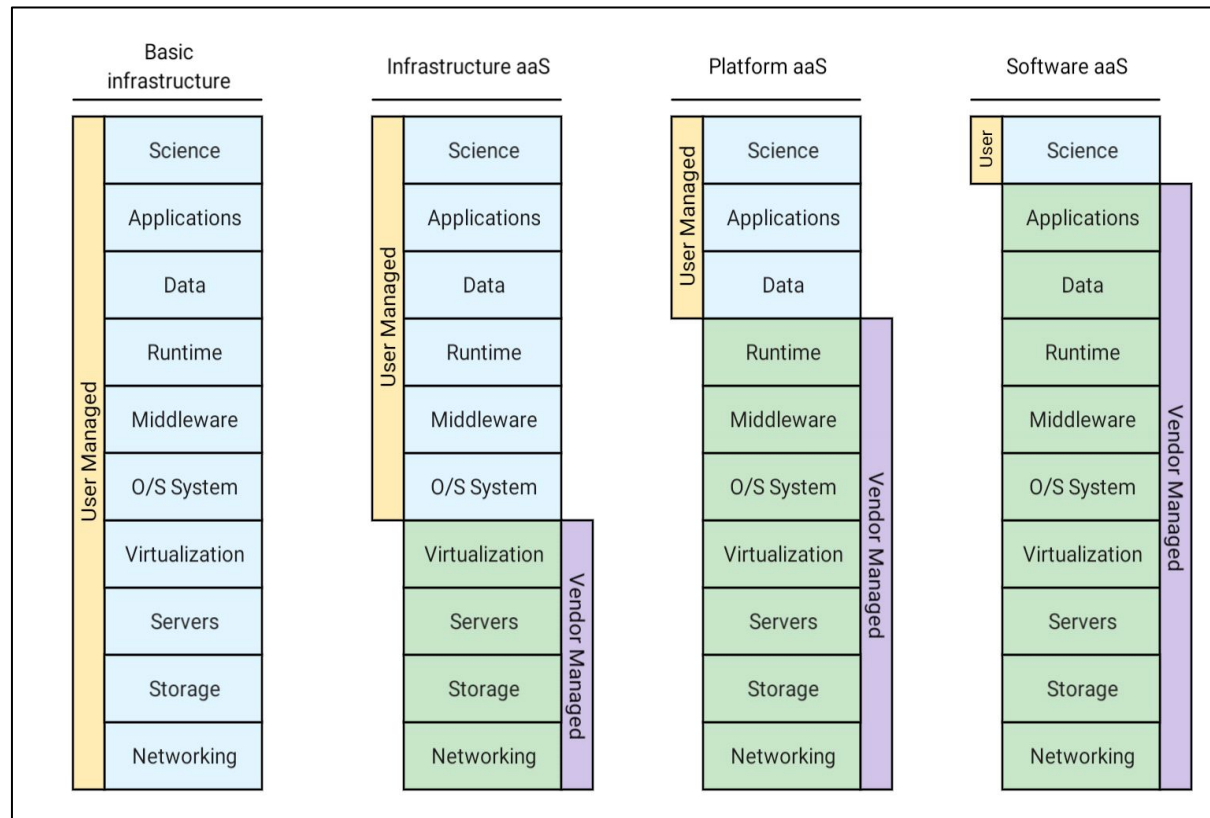


[^]arguable

*arguable (CI, CD)

What kind of science/projects? → Which model

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



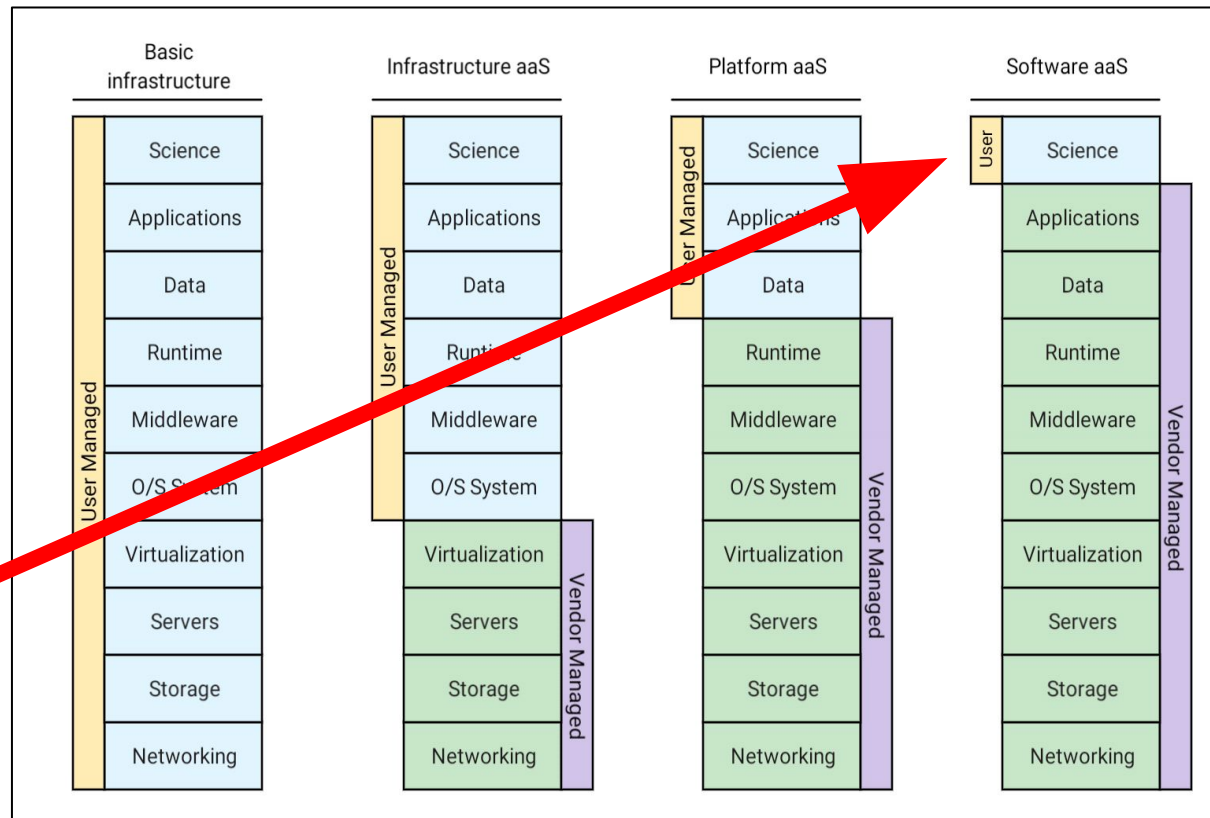
*arguable

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 (SClaaS?)

*arguable



Which Clouds?

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

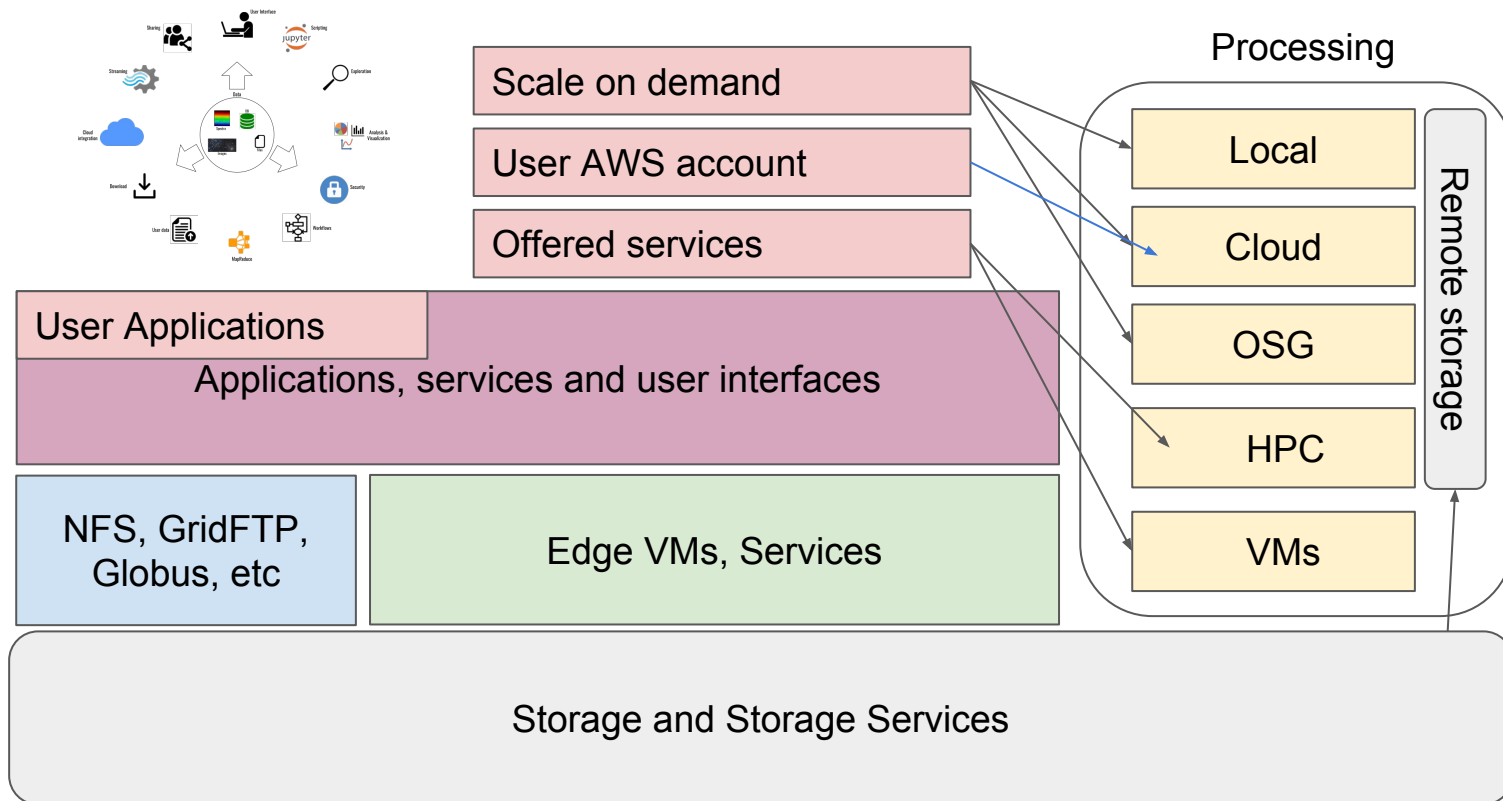
Salesforce, DigitalOcean, Rackspace,
 1&1, UpCloud, CityCloud, CloudSigma,
 CloudWatt, Aruba, CloudFerro, Orange,
 OVH, T-Systems



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

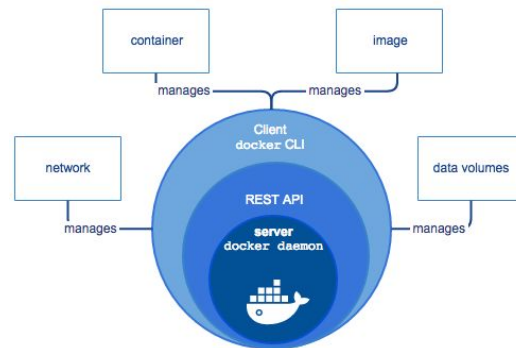


Why not both?



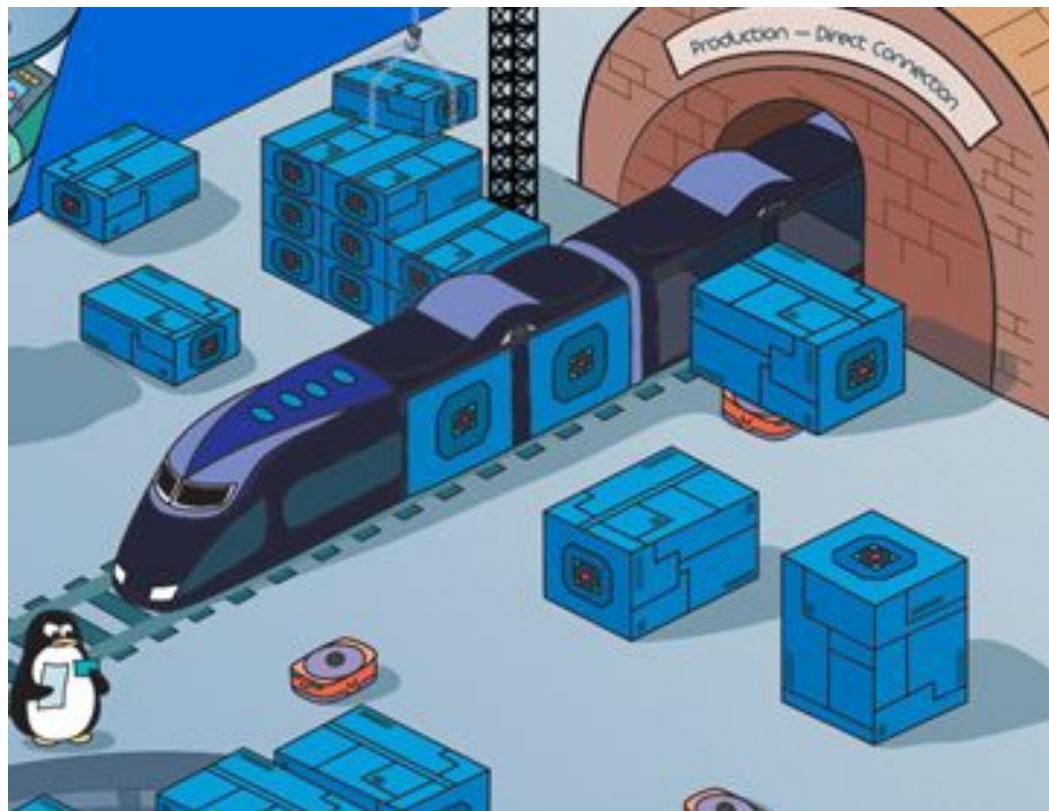
Containerization to the rescue

- 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, $\frac{1}{8}$ 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



```

DARK ENERGY SURVEY
DATA MANAGEMENT

easyaccess 1.4.0. The DESDM Database shell.
Connected as mcarras2 to desdct.
** Type 'help' or '?' to list commands. **

*General Commands* (type help <command>):
=====
clear edit help      history prefetch version
config exit help_function import shell

*DB Commands*      (type help <command>):
=====
add_comment      find_tables      myquota          show_index
append_table     find_tables_with_column mytables         user_tables
change_db        find_user         refresh_metadata_cache whoami
describe_table   load_table        set_password
execproc         loadsql           show_db

*Default Input*
=====
* To run SQL queries just add ; at the end of query
* To write to a file : select ... from ... where ... ; > filename
* Supported file formats (.csv, .tab., .fits, .h5)
* To check SQL syntax : select ... from ... where ... ; < check
* To see the Oracle execution plan : select ... from ... where ... ; < explain

* To access an online tutorial type: online_tutorial

DESDB ~>

```

- 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



DES Labs

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

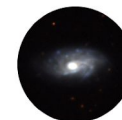
Easyaccess web



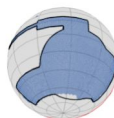
Jupyterhub + easyaccess



DES cutouts



Footprint



Easyaccess online



DESDM Services status



External Links

Science Server



NOAO Data Lab



CosmoHub



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...](#)

```
SELECT dr1.RA,dr1.DEC,dr1.COADD_OBJECT_ID
FROM dr1_main sample(8.01) dr1
WHERE
dr1.MAG_AUTO_G < 18 and
dr1.WAVG_SPREAD_MODEL_I + 3.0*dr1.WAVG_SPREADERR_M
dr1.WAVG_SPREAD_MODEL_I + 1.0*dr1.WAVG_SPREADERR_M
dr1.WAVG_SPREAD_MODEL_I - 1.0*dr1.WAVG_SPREADERR_M
dr1.WAVG_SPREAD_MODEL_I > -1 and
dr1.IMAFLAGS_ISO_G = 0 and
dr1.IMAFLAGS_ISO_R = 0 and
dr1.IMAFLAGS_ISO_I = 0 and
```

EXAMPLE QUERIES

See some example queries as a start

[More...](#)



CUTOUTS SERVICE

Generate cutouts for positions or ids

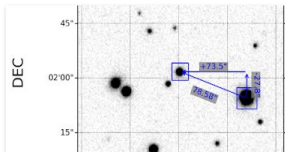
[More...](#)



DES JupyterLabs

(Beta) Jupyter Labs

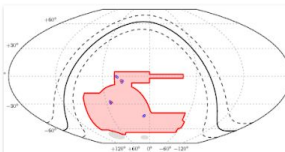
[More...](#)



FINDING CHART

Find your object

[More...](#)



DES FOOTPRINT

Interactive globe

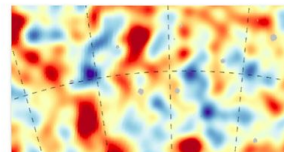
[More...](#)



DATA ANALYSIS

SEDs and color-color diagrams

[More...](#)



MY JOBS

List of submitted jobs

[More...](#)



HELP

Help form

[More...](#)

NCSA DESaccess: DB access

DARK ENERGY SURVEY desaccess



mck

mcarras2@illinois.edu

Query box

Insert your query in the box below. Data results for "Quick" Jobs (30 sec.) will be displayed at the bottom.

```

1 --
2 -- Example Query --
3 -- This query selects stars around the center of globular cluster M2
4 SELECT
5 COADD_OBJECT_ID,RA,DEC,
6 MAG_AUTO_G,G,
7 MAG_AUTO_R,R,
8 WAVG_MAG_PSF_G_PSF,
9 WAVG_MAG_PSF_R_PSF
10 FROM DR1_MAIN
11 WHERE
12 RA between 323.36-0.12 and 323.36+0.12 and
13 DEC between -0.82-0.12 and -0.82+0.12 and
14 WAVG_SPREAD_MODEL_I + 3.0*WAVG_SPREADERR_MODEL_I < 0.005 and
15 WAVG_SPREAD_MODEL_I > -1 and
16 IMAFLAGS_ISO_G = 0 and
17 IMAFLAGS_ISO_R = 0 and
18 FLAGS_G < 4 and
19 FLAGS_R < 4
20

```

Submit Job

Clear

Check

Quick

See Examples

Output file (.csv, .fits or .h5). Enable in order to submit.

Output file

Options:

Compressed files (csv and h5 files). Slightly longer jobs but smaller files

Job Name (optional)

Send email after completion

Email

Home

DB access

DR1 Table Schema

Example Queries

Cutout Service

DR1 Footprint


My Jobs

DES JupyterLab

Help

NCSA DESaccess: Cutouts Service

DARK ENERGY SURVEY desaccess
👤



mck
mcarras2@illinois.edu

- Home
- DB access
- DR1 Table Schema
- Example Queries
- Cutout Service
- DR1 Footprint
- My Jobs
- DES JupyterLab
- Help

Coads Images Cutout Form

Upload the file with the positions or enter the positions by hand and run the desthumb generator

- 📁 Upload File (csv, with RA,DEC as uncommented header)
- 📄 Enter Values
- 📏 Xsize (in arcminutes): 1.0
- 📏 Ysize (in arcminutes): 1.0
- ✍️ Job Name
- ✉️ Email Options
- 📁 Return Type

🗑️ Clear Form

📁 Upload File

📄 Enter Values

1


1

Send email on completion Email

Return just list of files (do not produce and display pngs, i.e. faster)

🚀 Submit Job


NCSA DESaccess: Cutouts Service



DARK ENERGY SURVEY desaccess

Job1 : d927a264_746c_4f7a_82cd_f46ebce496c7 (19 objects)

See Log
Download All
Close



mck
mcarras2@illinois.edu

- Home
- DB access
- DR1 Table Schema
- Example Queries
- Cutout Service
- DR1 Footprint
- My Jobs
- DES JupyterLab
- Help

Coadds Images Cutout Form

Upload the file with the positions or enter the positions by hand and run the desthumb!

📁 Upload File (csv, with RA,DEC as uncommented header)

📄 Enter Values

📏 Xsize (in arcminutes): 1.0

📏 Ysize (in arcminutes): 1.0

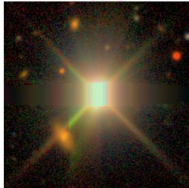
✍️ Job Name

✉️ Email Options

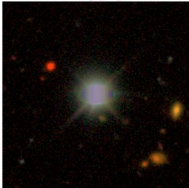
📄 Return Type

⚙️ Clear Form

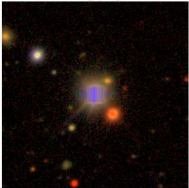
143782572_irg.png



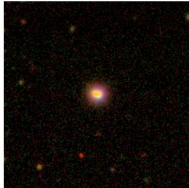
143830746_irg.png



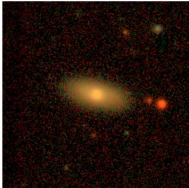
143821280_irg.png



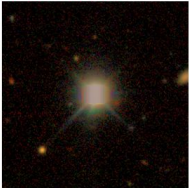
197515917_irg.png




154564429_irg.png



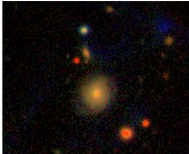
206021039_irg.png



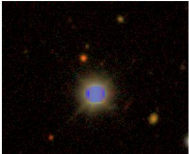
179484171_irg.png



149989928_irg.png



143470014_irg.png



👤

?

 1

 1

NCSA DESaccess: Asynchronous Jobs

DARK ENERGY SURVEY desaccess



mck
mcarras2@illinois.edu

My Jobs

#	Status	Job Name	Job type	Execution time (s)	Cancel Job	Queries	Results	Files
0	●	Name: Job id: 6b4cac2b-b544-44e1-96bf-58cd4968a338 6 days and 0 hours ago (Expired)	query	0	⊗	Query	Cutouts	Files
1	●	Name: Job id: daf5e3c-461e-42ed-8efb-5fcbf684047 6 days and 0 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
2	●	Name: testapi Job id: 0dfc5a58-b00a-4798-834f-9816c6f9a985 7 days and 4 hours ago (Expired)	cutout	3	⊗	Query	Cutouts	Files
3	●	Name: testapi Job id: 12961656-8075-4629-8e4f-fd4378013634 7 days and 4 hours ago (Expired)	cutout	3	⊗	Query	Cutouts	Files
4	●	Name: testapi Job id: 09a37f69-2096-4296-b87d-c6567cde0649 7 days and 4 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
5	●	Name: Job id: fd10cf32-3cc6-4090-bb90-344268dd615e 7 days and 5 hours ago (Expired)	cutout	1	⊗	Query	Cutouts	Files
6	●	Name: testapi Job id: b85ea747-5201-4e49-a0eb-f2b667f266de 7 days and 5 hours ago (Expired)	cutout	-1	⊗	Query	Cutouts	Files
7	●	Name: Job id: 8f8fa56a-4685-49f9-b7be-603310ccdddb 8 days and 16 hours ago (Expired)	query	577	⊗	Query	Cutouts	Files
8	●	Name: Job id: df8a57c4-b1d5-4332-80d5-a08a27b537d9 8 days and 16 hours ago (Expired)	query	1042	⊗	Query	Cutouts	Files
9	●	Name: Job id: 7f1db550-4d38-441f-a037-ed659b3b79c9 8 days and 16 hours ago (Expired)	query	-1	⊗	Query	Cutouts	Files
10	●	Name: Job id: fcaacde0-9d63-45d4-92f2-4f847b9b415c 8 days and 16 hours ago (Expired)	query	9	⊗	Query	Cutouts	Files
11	●	Name: Job id: a88b79cc-fd71-4ee0-a33d-92b5be98106f 8 days and 17 hours ago (Expired)	query	9	⊗	Query	Cutouts	Files
	●	Name: demo1						


REFRESH ↻

DELETE 🗑

- Home
- DB access
- DR1 Table Schema
- Example Queries
- Cutout Service
- DR1 Footprint
- My Jobs**
- DES JupyterLab
- Help

NCSA DESaccess: Footprint and Jupyter Labs

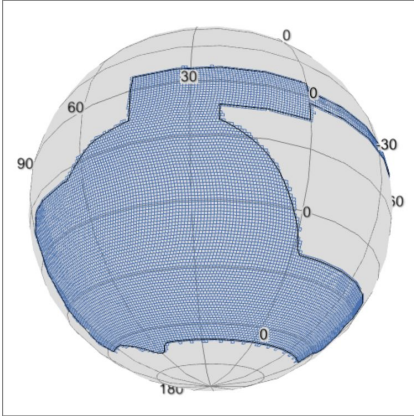
I DARK ENERGY SURVEY desaccess



mck
mcarras2@illinois.edu

- Home
- DB access
- DR1 Table Schema
- Example Queries
- Cutout Service
- DR1 Footprint**
- My Jobs
- DES JupyterLab
- Help

DES DR1 Footprint



Use the footprint tool to search a tile by position or name. Double click to select a tile.

Position (ra,dec) Tile name

Coordinates
 DR1 TILES
 HPIX nside=32


Tile properties

Name :
 Tile Center :
 No Objects :
 RA Corners :
 DEC Corners :

Get Tile Files

Click [here](#) to get access to all the tiles

I DARK ENERGY SURVEY desaccess



mck
mcarras2@illinois.edu

- Home
- DB access
- DR1 Table Schema
- Example Queries
- Cutout Service
- DR1 Footprint
- My Jobs
- DES JupyterLab**
- Help

DES Jupyter Labs (Beta)

This feature is experimental only. Please use with caution. You can launch, access and delete your Jupyter Notebook. This Notebook will run with 1 CPU and 2GB of RAM.

Deploy Lab +
Delete Lab

Status

● Ready

Status: Running

Go To Lab ↕

REFRESH ↻

NCSA DESaccess: Labs with access to Jobs and easyaccess

The screenshot displays a JupyterLab environment with the following components:

- File Browser:** Shows a directory structure under 'jobs' with various files and folders, including their last modified dates (e.g., '7 days ago', '10 hours ago').
- Scatter Plot:** A plot titled 'basics_plotting' showing a positive correlation between 'MAG_AUTO_R' (x-axis, 18-26) and 'MAG_AUTO_I' (y-axis, 18-26). The data points are overlaid with a density heatmap and green marginal distributions.
- Code Editor:** Contains Python code using Holoviews and Bokeh:


```
In [9]: import holoviews as hv
        hv.extension('bokeh')

In [10]: hextiles = hv.HexTiles(df, [('MAG_AUTO_R', 'R'), ('MAG_AUTO_I', 'I')], [], extents=(20,26,20

In [11]: hextiles.options(width=500, height=500, min_count=0, tools=['hover'], colorbar=True, ) * hv.

Out[11]:
```
- Terminal:** Shows the 'easyaccess' shell interface for the DESDB database. It includes a hexagonal logo and the text 'DARK ENERGY SURVEY DATA MANAGEMENT'. The terminal output shows:


```
easyaccess 1.4.4. The DESDB Database shell.
Connected as mck to desdb.
** Type 'help' or '?' to list commands. **

*General Commands* (type help <command>):
=====
clear edit help history prefetch version
config exit help_function import shell

*DB Commands* (type help <command>):
=====
describe_table loadsql show_db
find_tables refresh_metadata_cache show_index
find_tables_with_column set_password whoami

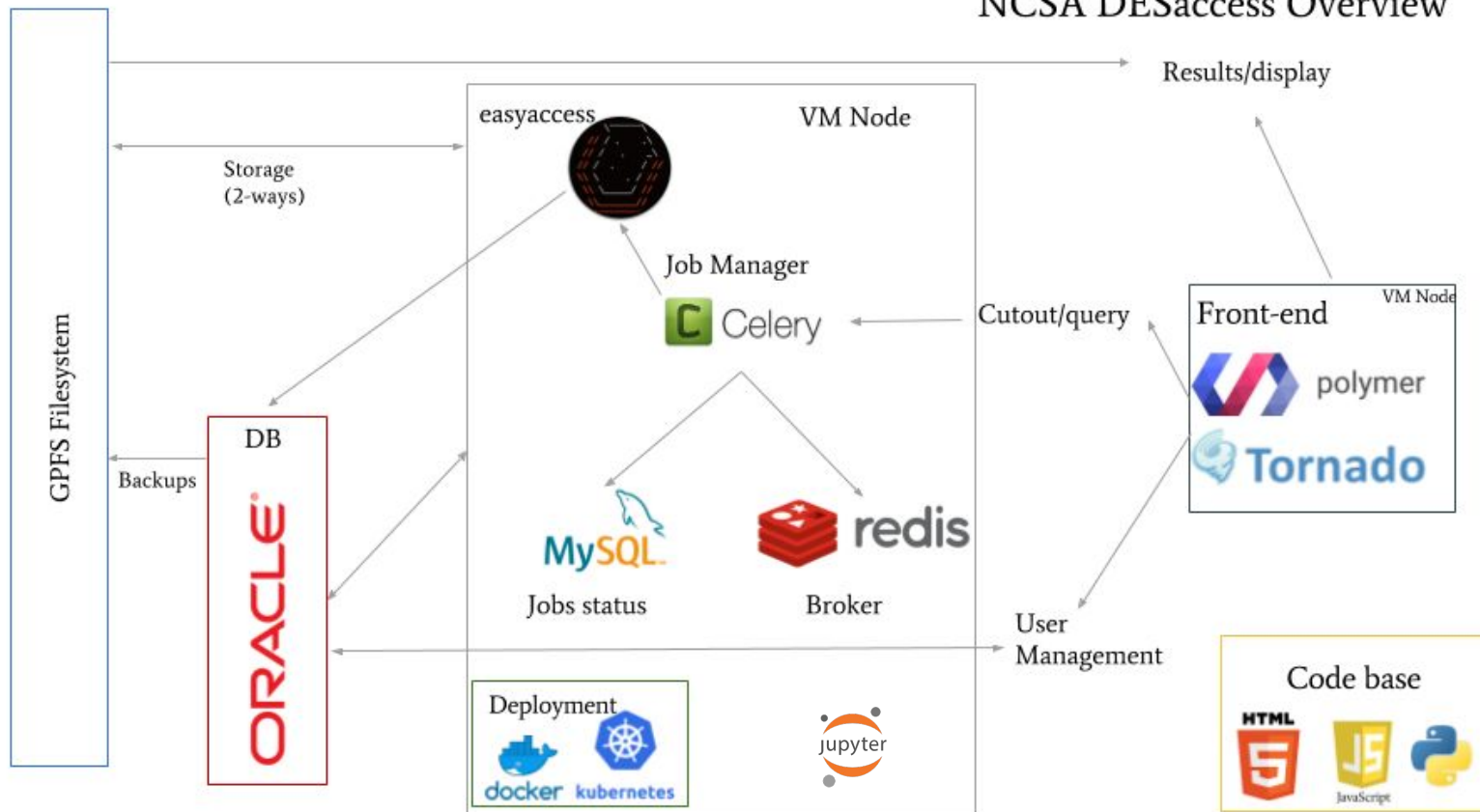
*Default Input*
=====
* To run SQL queries just add ; at the end of query
* To write to a file : select ... from ... where ... ; > filename
* Supported file formats (.csv, .tab, .fits, .h5)
* To check SQL syntax : select ... from ... where ... ; < check
* To see the Oracle execution plan : select ... from ... where ... ; < explain

* To access an online tutorial type: online_tutorial

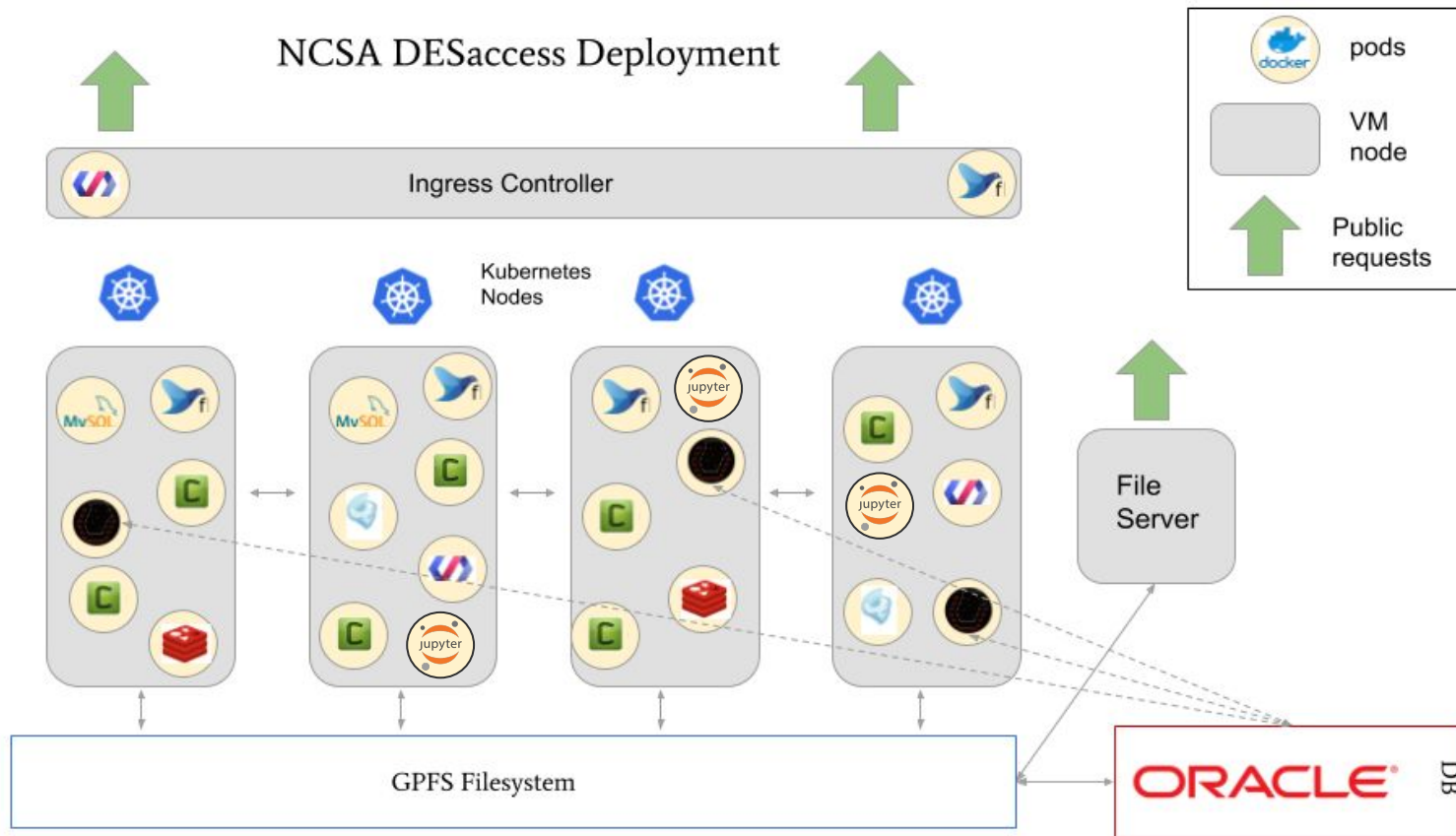
DESDB -> |
```

NCSA DESaccess: Technology Overview

NCSA DESaccess Overview



NCSA DESaccess: Deployment



LSST Science Platform



LSST Users

Internet

LSST Science Platform



Portal

JupyterLab



Web APIs



Data Releases



Alert Streams



User Databases



User Files



User Computing



Software Tools

gaia archive

[HOME](#) [SEARCH](#) [STATISTICS](#) [VISUALISATION](#) [DOCUMENTATION](#) [HELP](#)

Welcome to the Gaia Archive

Gaia is an ambitious mission to chart a three-dimensional map of our Galaxy, the Milky Way, in the process revealing the composition, formation and evolution of the Galaxy. Gaia will provide unprecedented positional and radial velocity measurements with the accuracies needed to produce a stereoscopic and kinematic census of about one billion stars in our Galaxy and throughout the Local Group. This amounts to about 1 per cent of the Galactic stellar population.



Top Features



Citation

How to cite and acknowledge Gaia.



Search

Query for Gaia sources using an ADQL (Astronomical Data Query Language) interface in an asynchronous mode (UWS).



Download

Direct download of Gaia data files.



Help

For questions, suggestions or problem reports, contact the Helpdesk.



Documents

Links to Gaia Archive and related Gaia documentation.



Gaia Mission

News, information, and resources on the Gaia mission for the scientific community.



Statistics

Show statistics of Gaia tables.



Partners

Partner data centres also serving Gaia data.

SciServer

Collaborative data-driven science

new message
[JIRA] (LSST-1
cluster to 1.11

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



SciServer is funded by National Science
Foundation award ACI-1261715



I 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

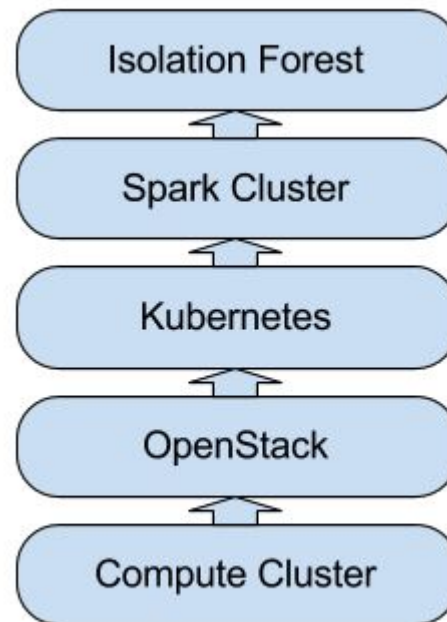
Technology Stack For Anomaly Service

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

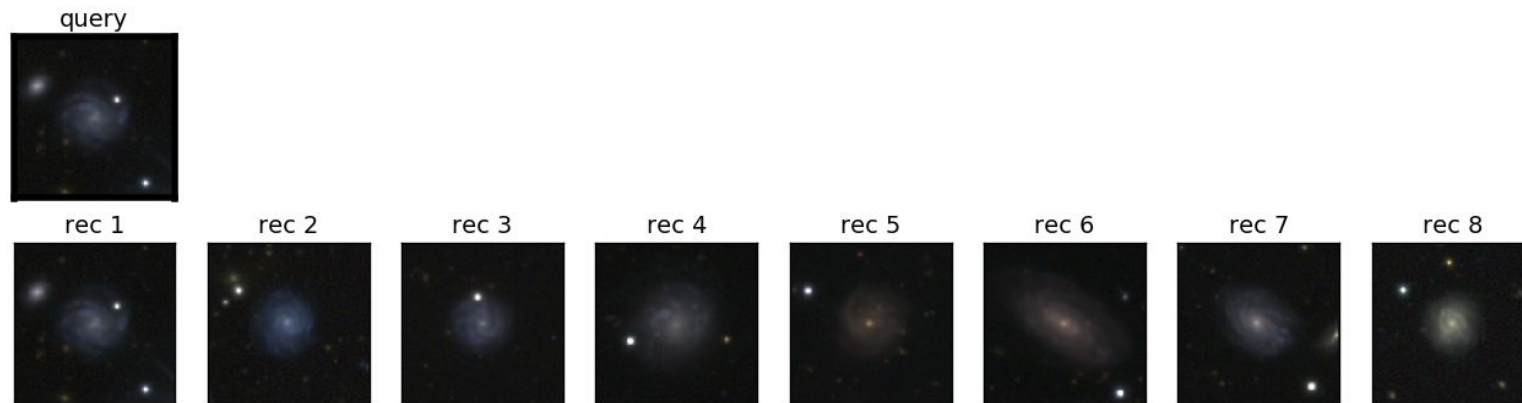
Sahand Hariri*
University of Illinois at Urbana-Champaign
sahandha@gmail.com

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

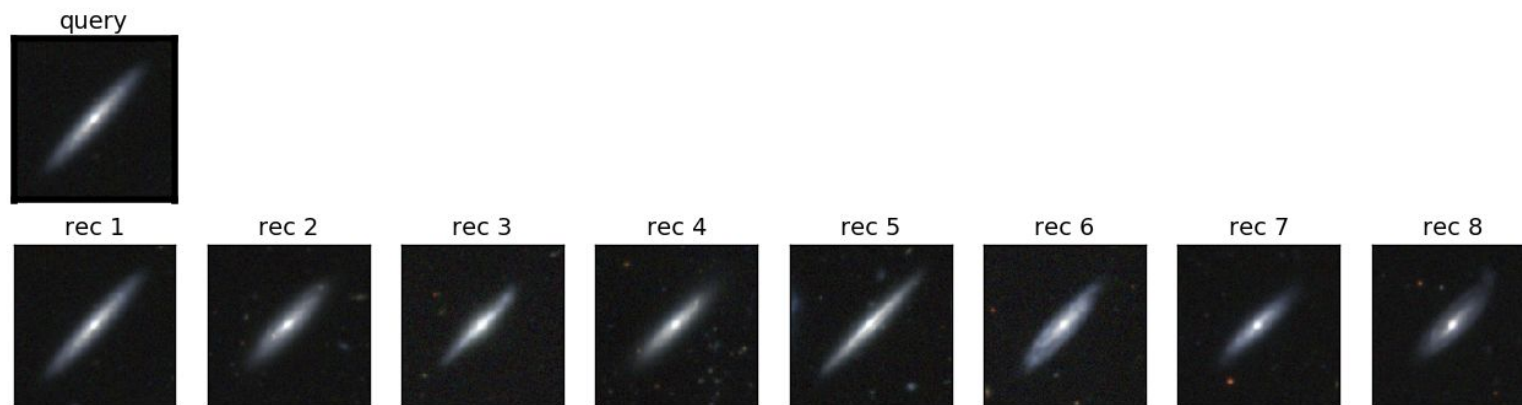
- Use Extended Isolation Forest as core algorithm
- Use Spark to parallelize trees and scoring
- Use Redis as a broker communicator
- To easily deploy in any environment, use Docker
- For orchestration of Docker containers, use Kubernetes
- Kubernetes cluster built on top of OpenStack, but it can be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



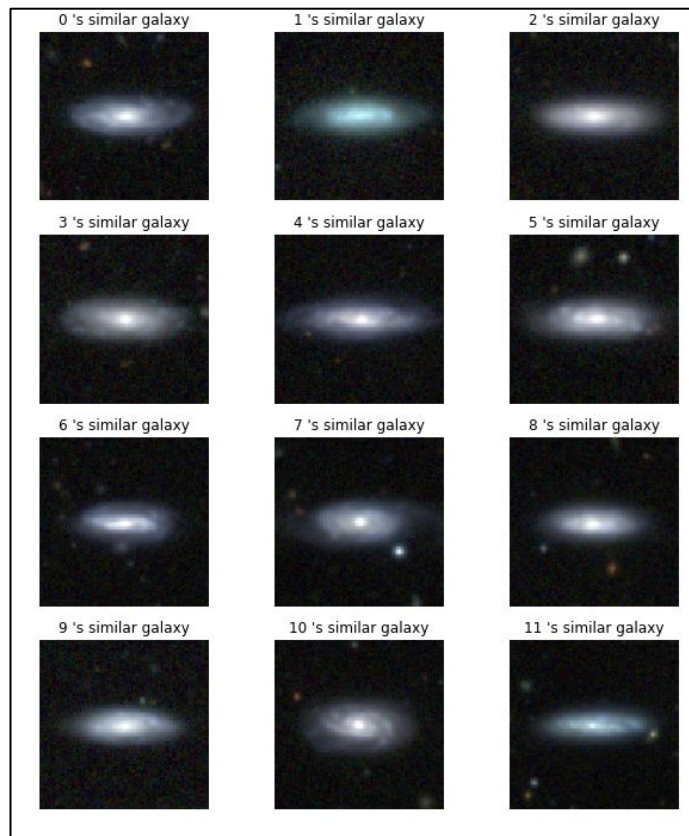
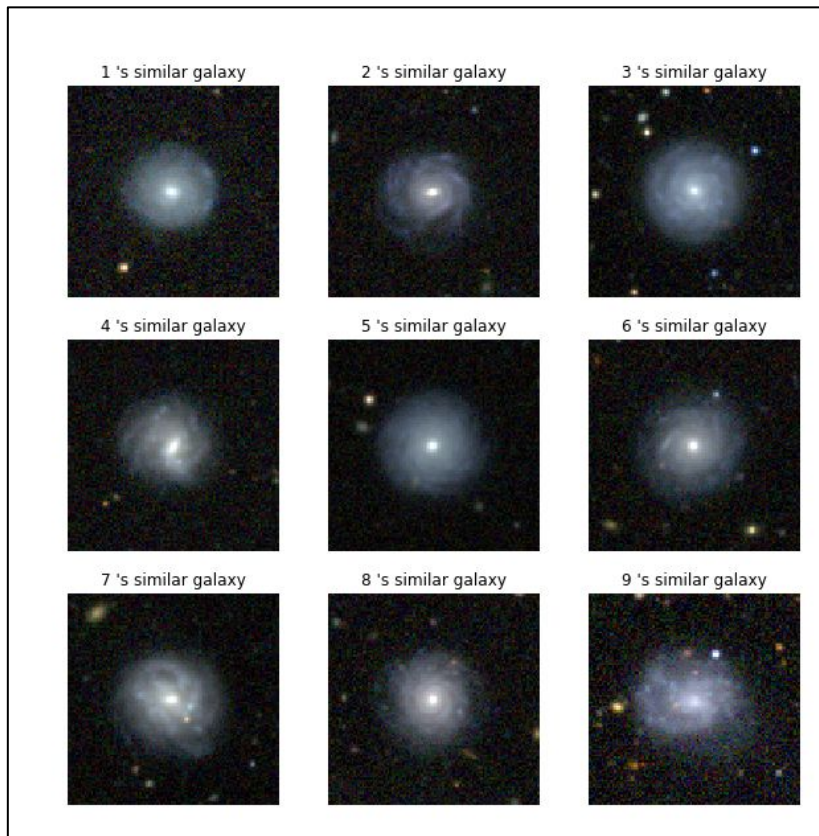
SClaaS Example: Galaxy selection and similarity search



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



SClaaS Example: Galaxy selection and similarity search



Final Remarks

Matias Carrasco Kind -- NCSA

mcarras2@illinois.edu

github.com/mgckind

matias-ck.com

- 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

Thank you!

Questions?

Matias Carrasco Kind -- NCSA

mcarras2@illinois.edu

github.com/mgkind

matias-ck.com