From SDSS to GenAI: retooling SciServer for the age of Foundation Models

(trustworthy, robust, efficient, interpretable)

Gerard Lemson

Institute for Data Intensive Engineering and Science
Data Science and AI institute
Johns Hopkins University

Johns Hopkins University | Data Science and AI Institute

JOHNS HOPKINS UNIVERSITY | EST. 1876 | JH150 | AMERICA'S FIRST RESEARCH UNIVERSITY

JOHNS HOPKINS UNIVERSITY

Data Science and AI Institute

Info For | Contact Us | Search

HOME

Data Science and AI Trusted Dataset Awards

DSAI Institute

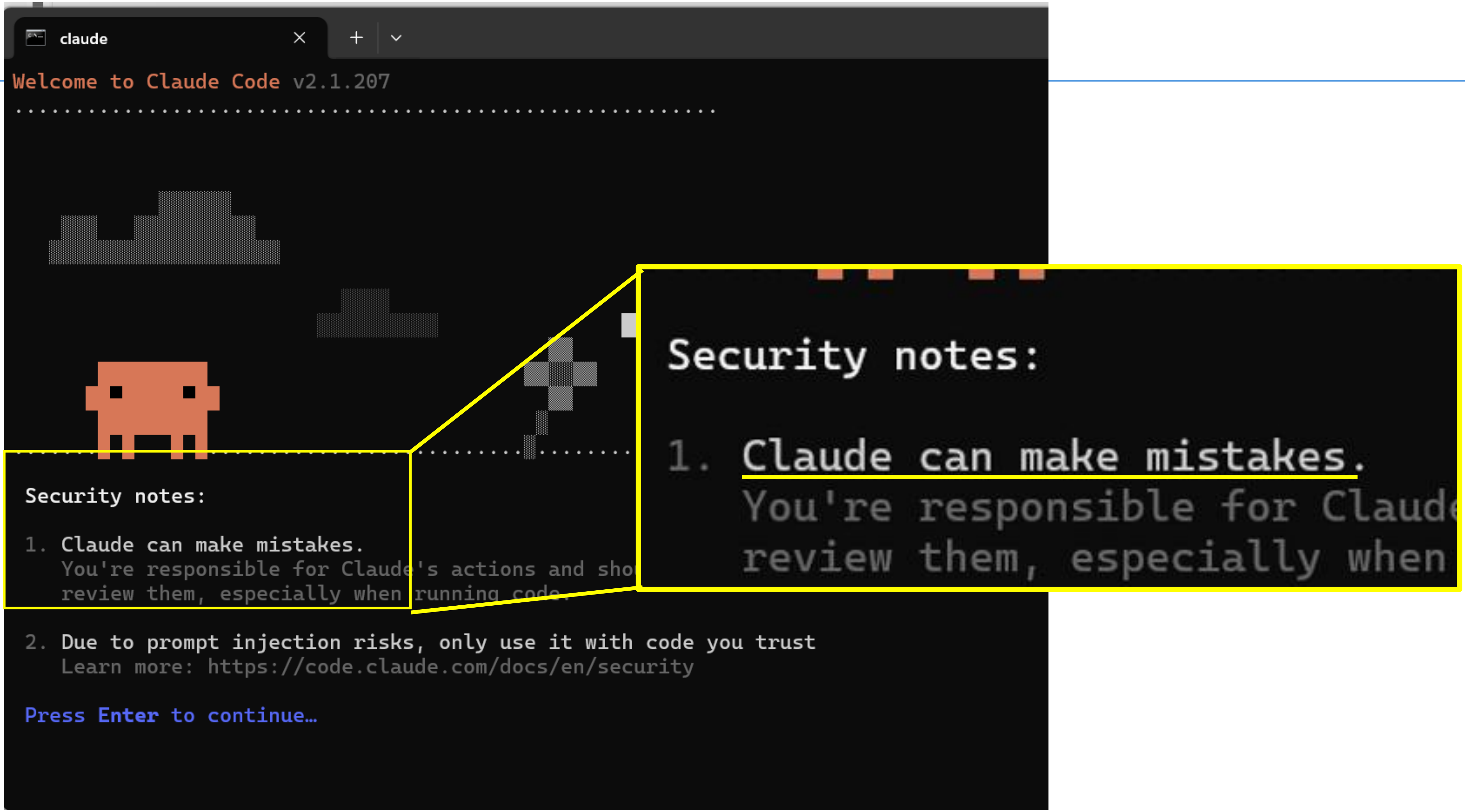
- 100+ faculty
  - BDP clusters
- >30 RSEs
  - Follow https://ai.jhu.edu/careers/
- New 500000 sqft building

NAIRR:

there is a need for *"Trusted data providers and hosts for a transparent and responsible AI data commons. Access to data should be tiered, controlled by the data providers, and provided through the same portal through which computational resources are provided."*

JOHNS HOPKINS UNIVERSITY | Data Science and AI Institute

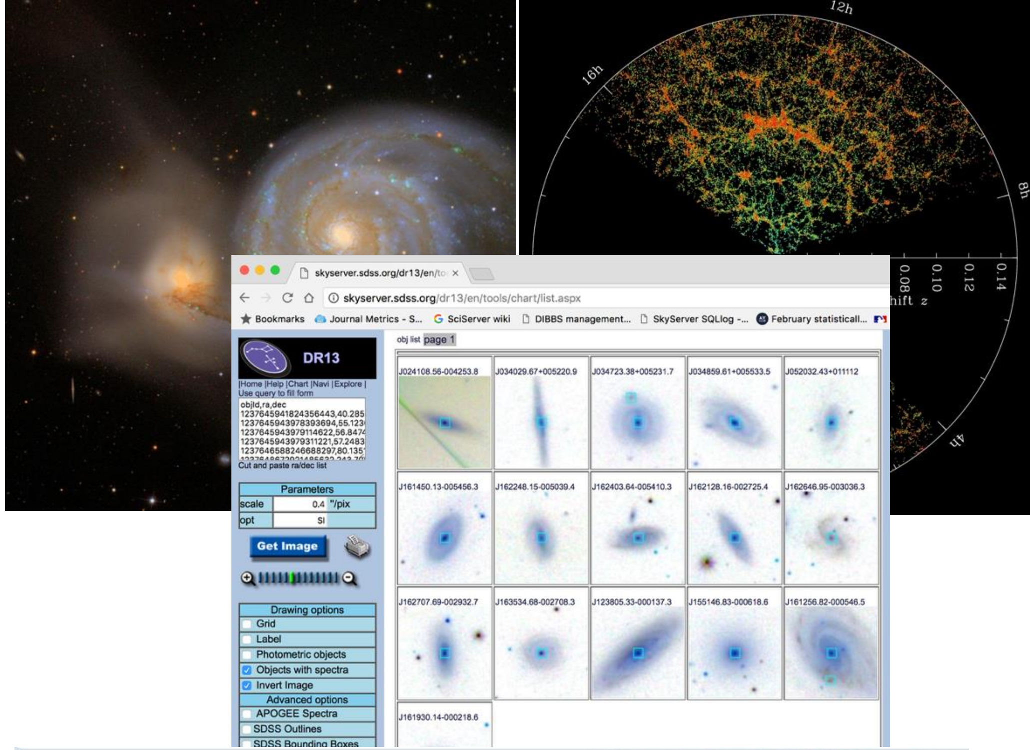
idies The Institute for Data-Intensive Engineering and Science

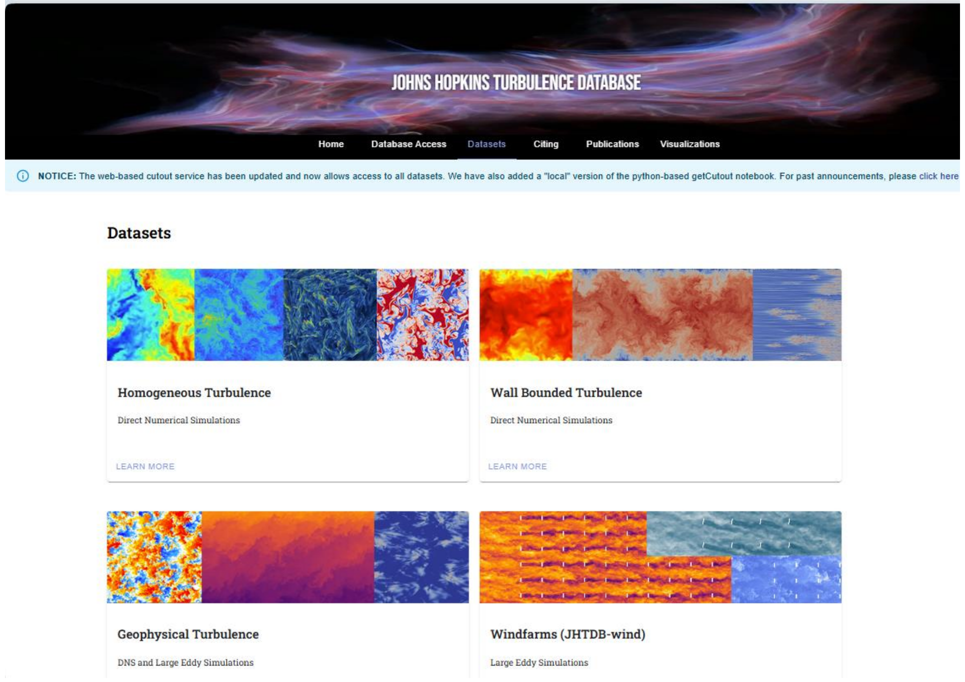

claude
Welcome to Claude Code v2.1.207
Security notes:
1. Claude can make mistakes.
You're responsible for Claude's actions and sho
review them, especially when running code.
2. Due to prompt injection risks, only use it with code you trust
Learn more: https://code.claude.com/docs/en/security
Press Enter to continue…
Security notes:
1. Claude can make mistakes.
You're responsible for Claude
review them, especially when

IDIES's history of creating trust

- Started with skyserver.sdss.org (2001)
  - Sloan Digital Sky Survey
- <u>Goal</u>: **instant access** to rich content
- <u>Idea</u>: **bring the analysis to the data**
- <u>Trust</u>: **Interactive access** at the core

- Extended to JHU Turbulence Database
  - **737,593,886,025,704 points queried**



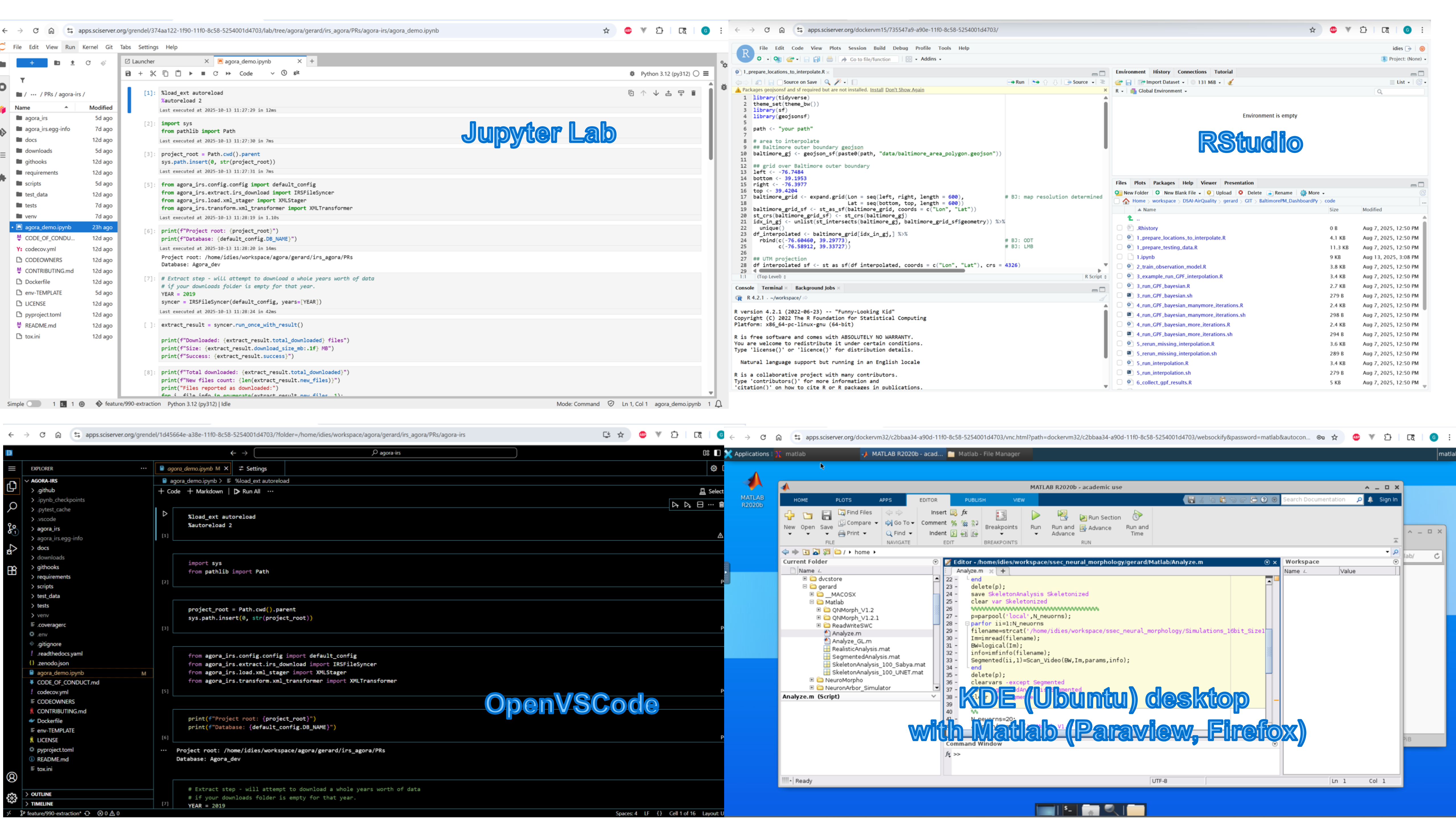


From SkyServer to SciServer

SciServer @ IDIES: www.sciserver.org

Supporting data intensive science everywhere

Data growing exponentially	Data too large for local copies: data-proximate analysis
Data becoming increasingly open/public	Difficult to aggregate large, geographically distributed data sets: joint analysis requires co-location
Data sets increasingly heterogeneous	Diverse set of skills and knowledge required, Joint analysis requires integration: aggregate data from various sources in a common context
Science increasingly collaborative	Data heterogeneity requires collaboration, sharing expertise
Analysis requiring increasingly sophisticated ML/AI tools	Sharing data, hardware, software, expertise


Jupyter Lab
RStudio
OpenVSCode
KDE (Ubuntu) desktop
with Matlab (Paraview, Firefox)

Astronomy: Database catalogues and images

jupyter 2. Stripe82-coadd-Copy1 Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Help | Python 3

CellToolbar

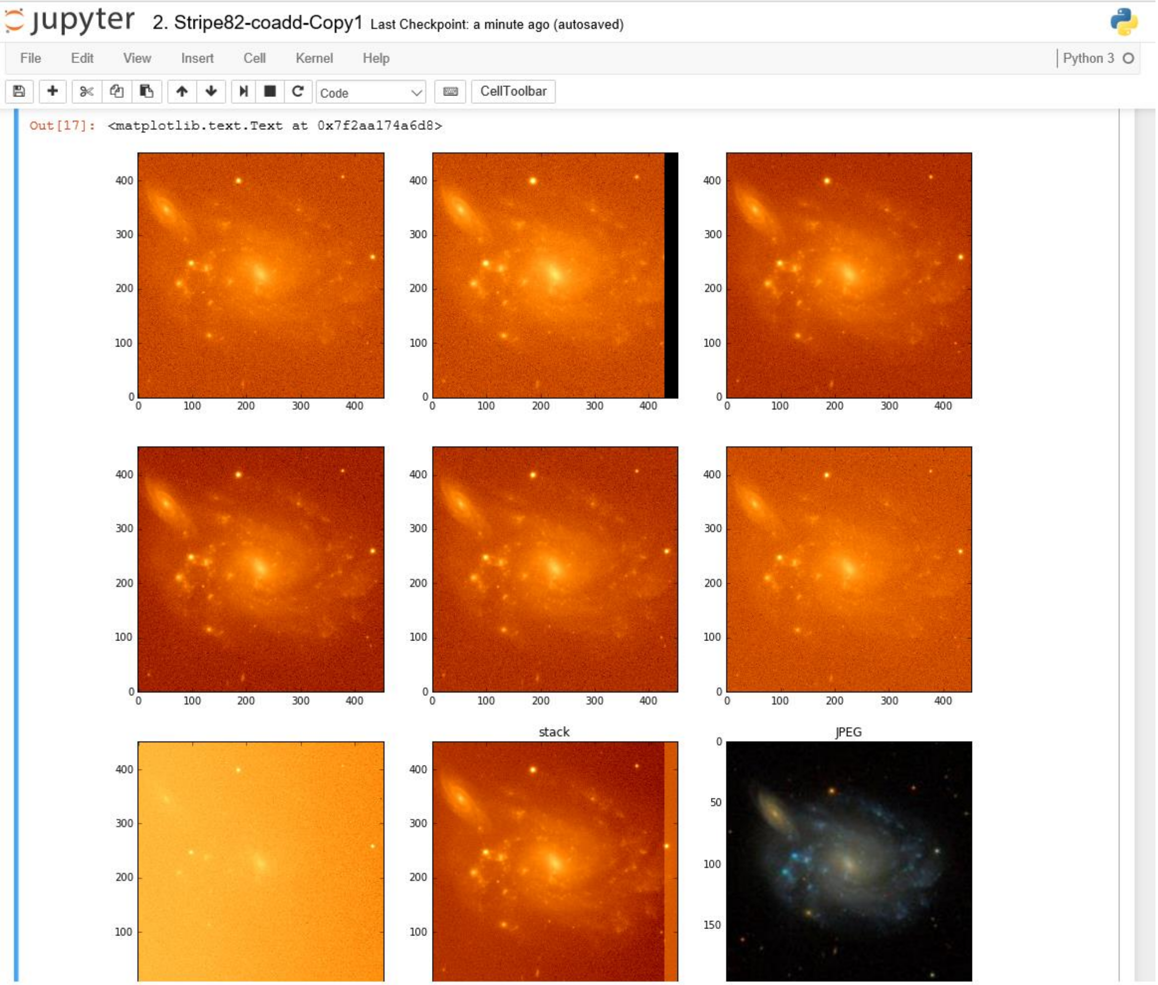
```python
import SciServer.LoginPortal as Login
token=Login.getToken()
import pandas
import tables
import numpy as np
import astropy
from astropy.io import fits
from astropy import wcs
import skimage.io
import urllib
import os
import matplotlib
import matplotlib.pyplot as plt
```

In [18]:

```python
sql="""
SELECT a.objid as head, c.objid2 as match, b.matchcount,
       p.fieldid as head_field, d.fieldid as match_field,
       dbo.fGetUrlFitsCFrame(d.fieldid, 'g') as fits_g,
       dbo.fGetUrlFitsCFrame(d.fieldid, 'r') as fits_r,
       dbo.fGetUrlFitsCFrame(d.fieldid, 'z') as fits_z,
       p.ra, d.ra as match_ra, p.dec, d.dec as match_dec
       , p.petror90_r
  from (select top 1 * from galaxy where objId=8658194378960928809) a
   join matchhead b on a.objid=b.objid        -- join with matchhead
   join photoobj p on a.objid=p.objid         -- get matchhead photoobj
   join match c on c.objid1=b.objid           -- join with all the matches
   join photoobjall d on c.objid2=d.objid     -- get match photoobj
order by d.fieldid                              -- order by matchhead objid
"""
queryResponse = SciServer.CasJobs.executeQuery(sql, "Stripe82",token=token)
obss = pandas.read_csv(queryResponse,index_col=None)
obss[:10]
```

Out[18]:

	head	match	matchcount	head_field	match_field	fits_g
0	8658194378960928809	8658194430499553320	57	8658194378960928768	8658194430499553280	http://das.sdss.org/imaging/5622/40/corr/
1	8658194378960928809	8658194477742948377	57	8658194378960928768	8658194477742948352	http://das.sdss.org/imaging/5633/40/corr/
2	8658194378960928809	8658194516375371821	57	8658194378960928768	8658194516375371776	http://das.sdss.org/imaging/5642/40/corr/
3	8658194378960928809	8658194585083510793	57	8658194378960928768	8658194585083510784	http://das.sdss.org/imaging/5658/40/corr/
4	8658194378960928809	8658194804163018771	57	8658194378960928768	8658194804163018752	http://das.sdss.org/imaging/5709/40/corr/
5	8658194378960928809	8658194954470752297	57	8658194378960928768	8658194954470752256	http://das.sdss.org/imaging/5744/40/corr/
6	8658194378960928809	8658195018907910161	57	8658194378960928768	8658195018907910144	http://das.sdss.org/imaging/5759/40/corr/
7	8658194378960928809	8658195044651040803	57	8658194378960928768	8658195044651040768	http://das.sdss.org/imaging/5765/40/corr/
8	8658194378960928809	8658195066151239724	57	8658194378960928768	8658195066151239680	http://das.sdss.org/imaging/5770/40/corr/
9	8658194378960928809	8658195113395748890	57	8658194378960928768	8658195113395748864	http://das.sdss.org/imaging/5781/40/corr/



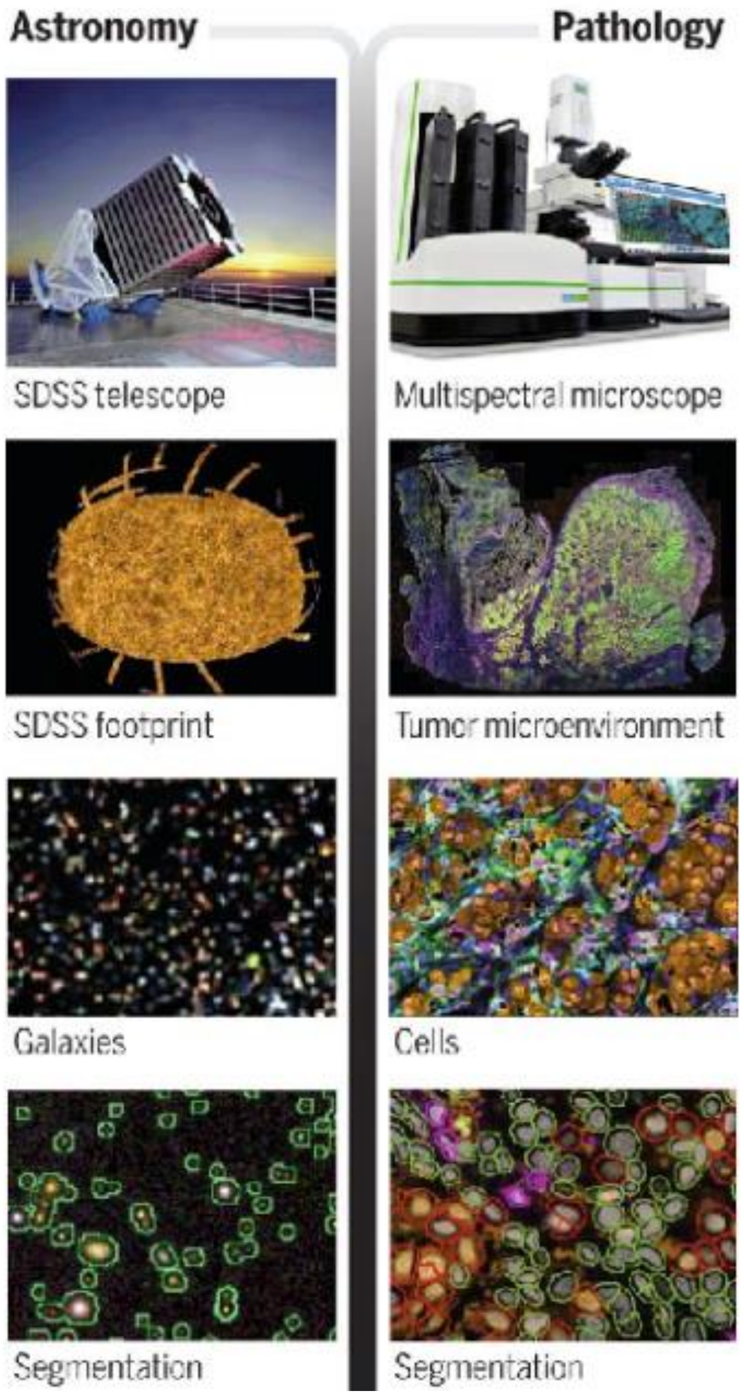
Johns Hopkins University | Data Science and AI Institute

idies The Institute for Data-Intensive Engineering and Science

AstroPath

Astronomy meets **Path**ology
Large-scale data management and analysis

Collaboration between IDIES and the Bloomberg-Kimmel Institute for Cancer Immunotherapy at JHH (Alex Szalay & Janis Taube)
Use high-precision analyses of specialized microscope imagery to learn about the tumor microenvironment, with the goal of impacting cancer immunotherapy:

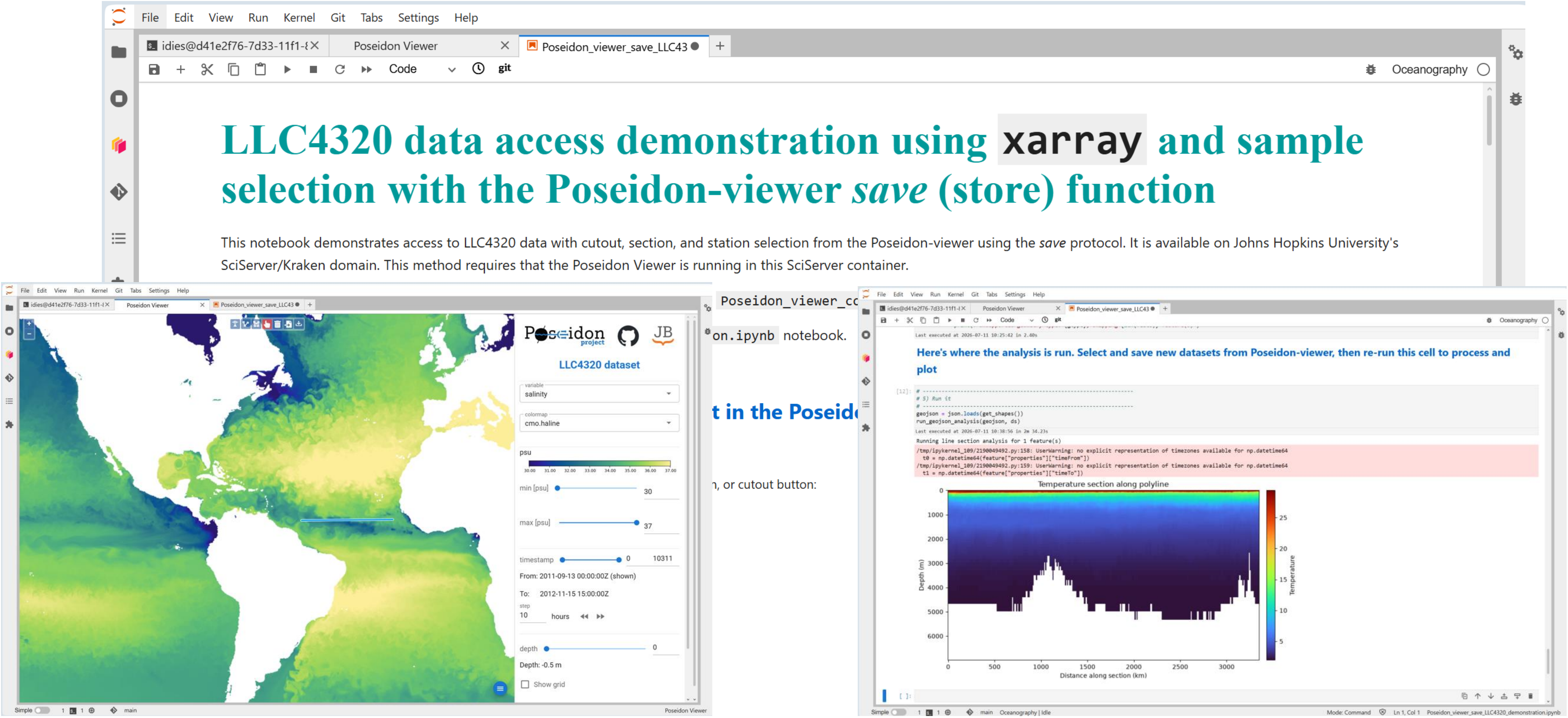


Credits Maggie Eminizer

Johns Hopkins University | Data Science and AI Institute

idies The Institute for Data-Intensive Engineering and Science

Poseidon ocean circulation simulations (4.5PB)



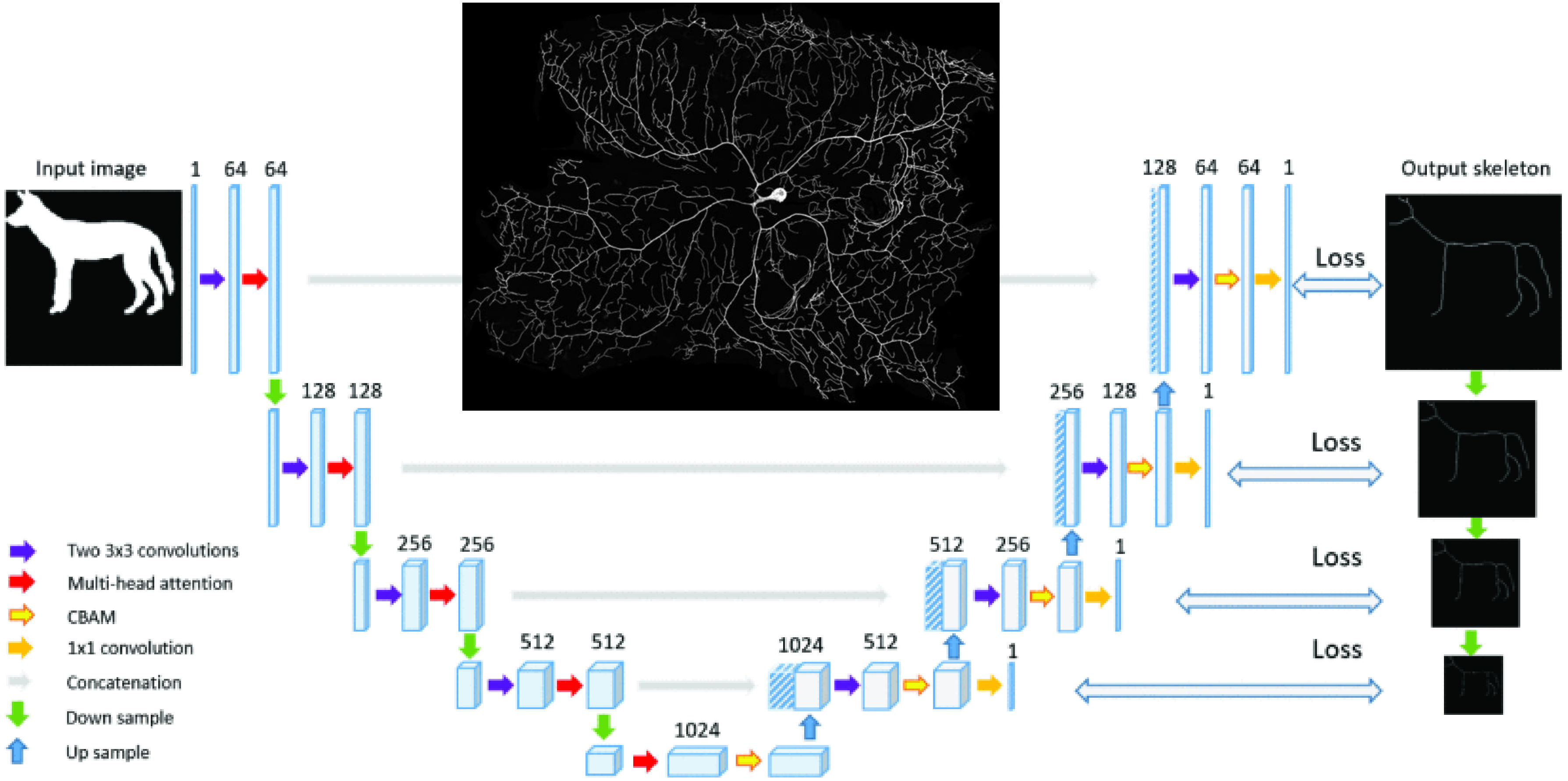
Johns Hopkins University | Data Science and AI Institute

idies
The Institute for Data-Intensive Engineering and Science

AI in SciServer: what can we support?

- Training?
  - We have images with tensorflow, pytorch
  - Upon request (risk of crypto mining) users can get access to compute environments with a100 GPUs
  - But SciServer is not an HPC system, we don't support massive training with 100 GPUs running for weeks just for you. Sorry.
- Inference?
  - Pre-trained models can be hosted and used for inference
    - Sometimes even on GPU
  - Embeddings in DB with vector support (postgres)
    - Similarity search (SpecCLIP example)
    - Clustering (SQL example)
- LLMs + GenAI
  - Text-to-SQL
  - Claude on SciServer
- AI ready, trusted datasets?
  - Creation multi-model datasets

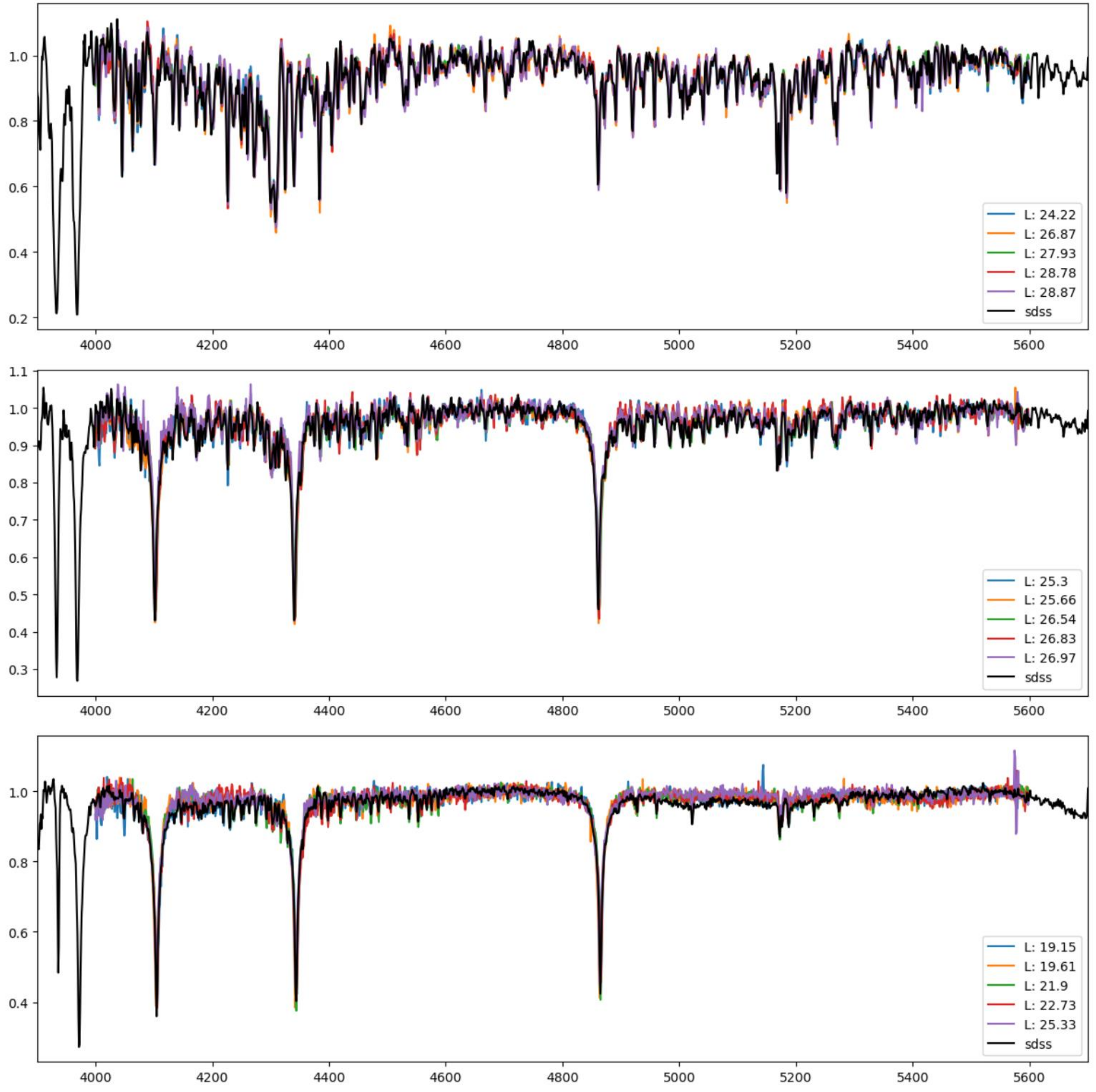
Training: U Net for image segmentation



SSEC: Adkins, Sutradhar (PIs), Vlad Kluzner, Ryan Hausen

SpecCLIP inference pipeline

- SpecCLIP created by Xiaosheng Zhao etal
  - Models loaded in shared SciServer data volume
  - >80k spectral embeddings (LAMOST, GAIA) stored in postgres vector DB
- Pipeline in a notebook
  - Query SDSS DR19 for stars for which there is a spectrum, include URL to spectrum file
  - Retrieve spectrum from local copy of SDSS SAS data
  - Create embedding the spectrum with SpecCLIP code provided by Xiaosheng
  - Use embedding as parameter to find 5 nearest LAMOST embeddings
  - Retrieve their spectra
  - Plot together with source spectrum

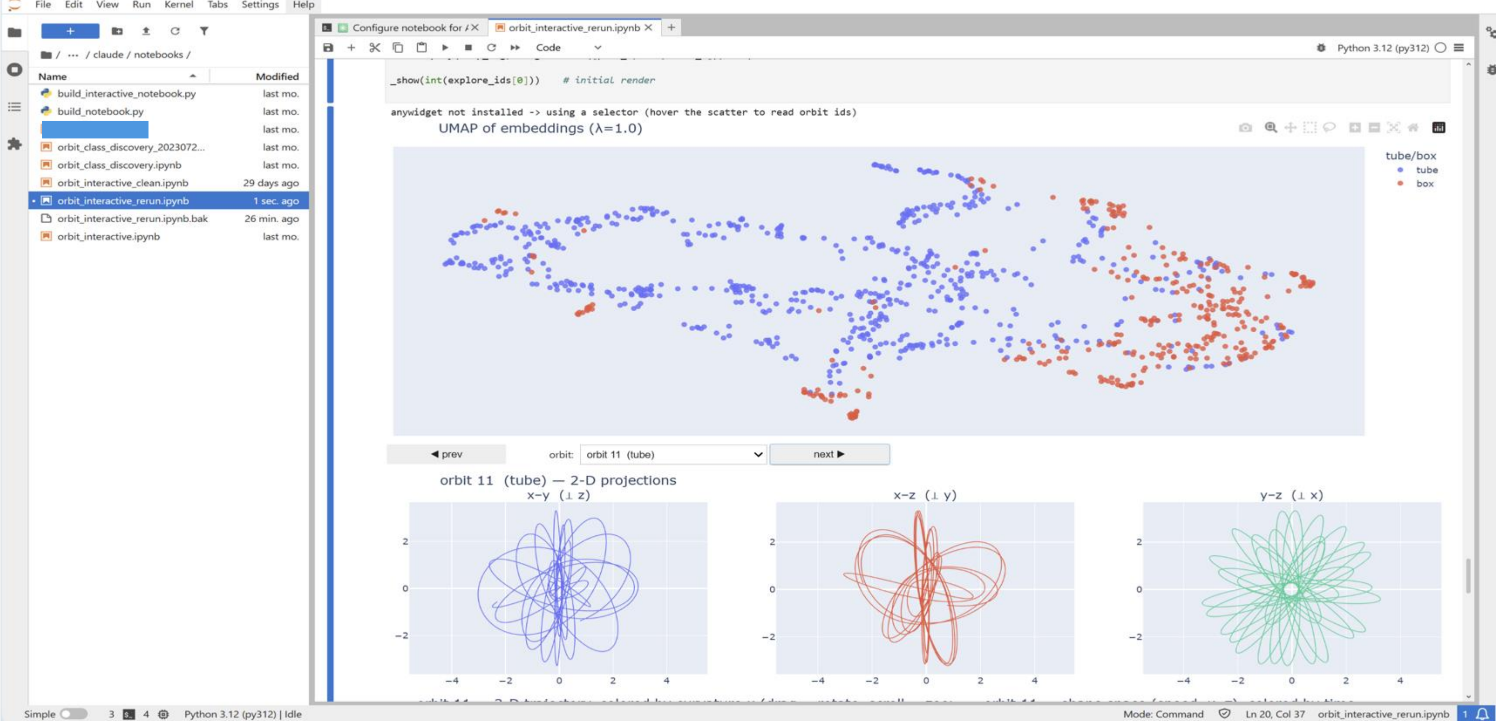

1.0
0.8
0.6
0.4
0.2
4000
4200
4400
4600
4800
5000
5200
5400
5600
L: 24.22
L: 26.87
L: 27.93
L: 28.78
L: 28.87
sdss
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
L: 25.3
L: 25.66
L: 26.54
L: 26.83
L: 26.97
sdss
L: 19.15
L: 19.61
L: 21.9
L: 22.73
L: 25.33
sdss

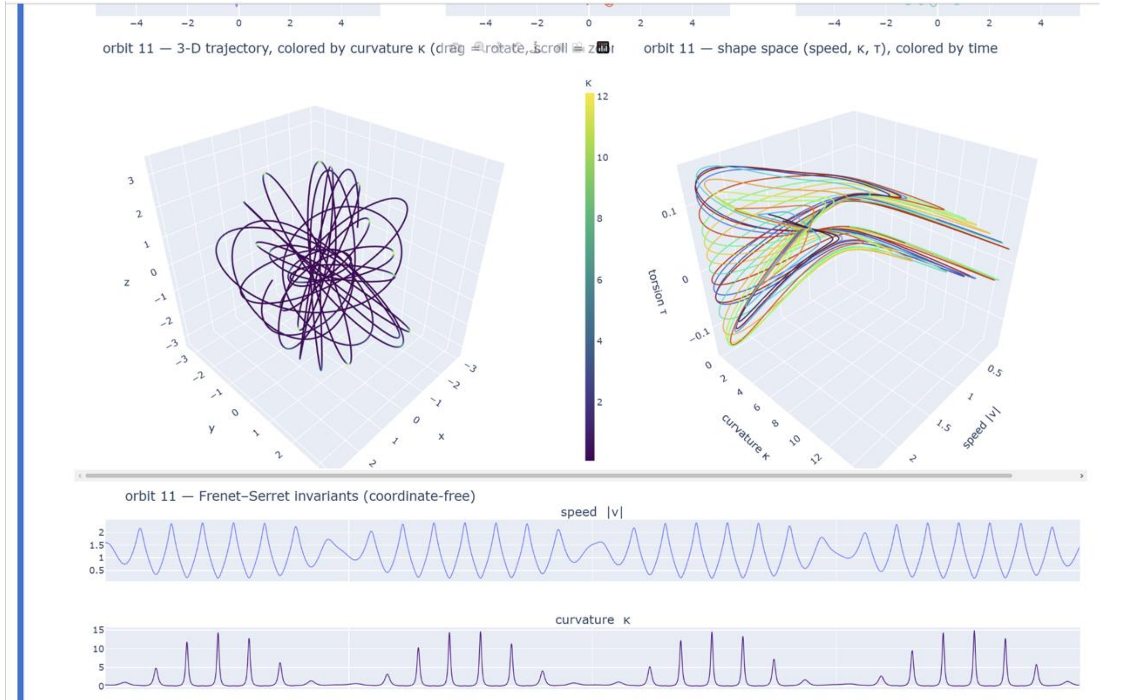
Data-proximate GenAI coding on SciServer

- Easy to install claude code and run it
  - Terminal or VS Code image
  - Can read and write data
  - Share sessions?
  - On your own dime
- Local chats with ollama or vllm possible
  - Some local models preloaded
  - We will explore setting up harnesses around them
  - Needs more work

Claude training on SciServer

- Asked claude to train a LeJEPA architecture on orbits in statistic potentials
- Orbits generated on SciServer with Nemo
- Create a notebook to plot projected (UMAP) embeddings.
- Active to clicks





LLMs and SciServer: Text-to-SQL

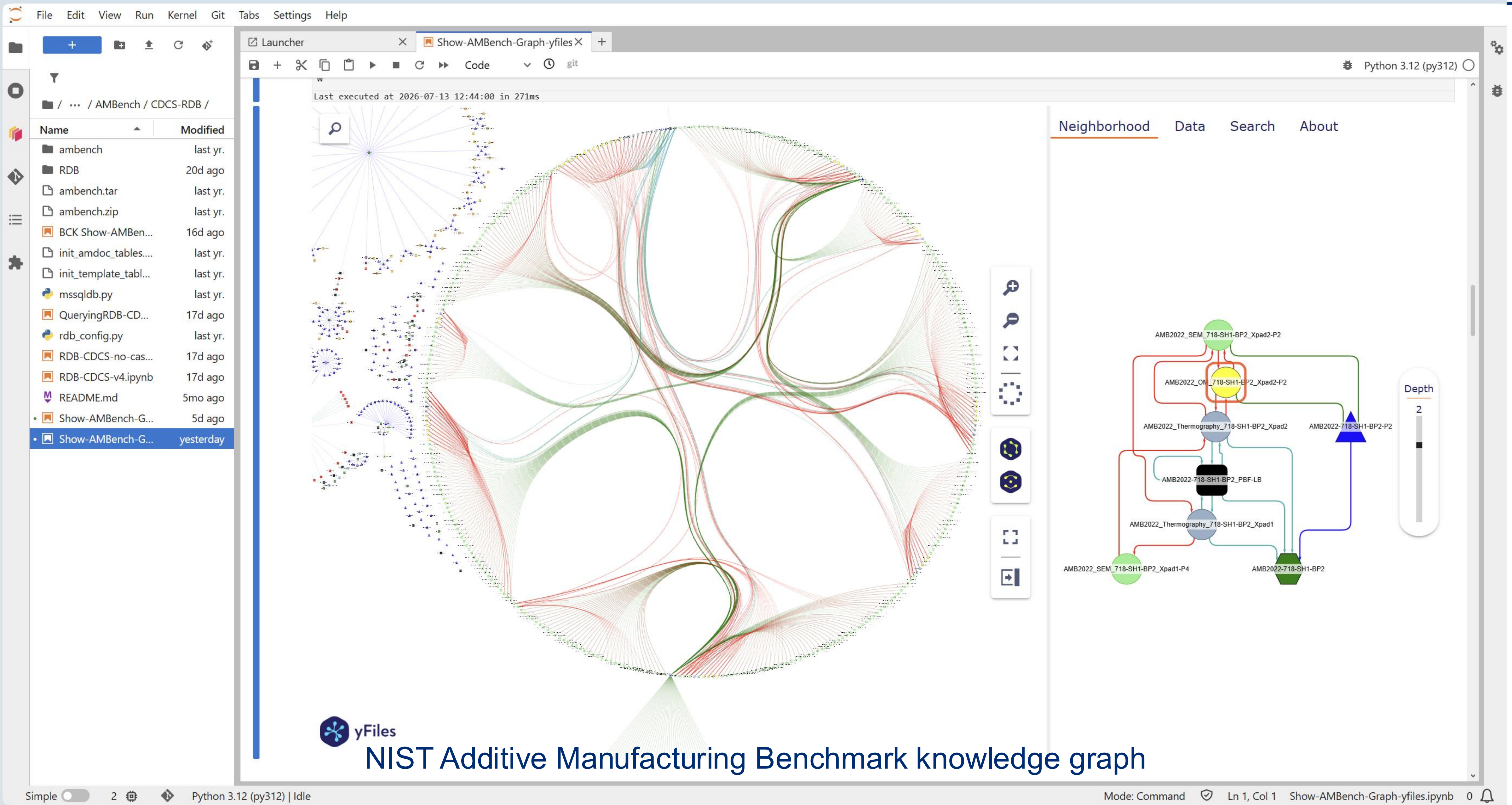
Can we give users a natural language (i.e. English for now) interface to query our relational databases

Intermezzo: Why use relational databases at all?

- encapsulation of data in terms of **standard** logical structure, *no need to know about internals of data storage*
- **standard** *query language* (SQL) for remote filtering
- advanced *query optimizers* (indexes, clustering, parallelism)
- authenticated remote access for multiple concurrent users

- **forces one to think carefully about data structure**
- **simplest way to support remote filtering**
- **speeds up path from science question to answer**
- **facilitates communication (query code is cleaner)**
- **custom views to simplify user queries**

Good for complex metadata



NIST Additive Manufacturing Benchmark knowledge graph

Text-to-SQL for SDSS using ChatGPT.com

Prompt: can you write a SQL query that queries the SDSS database for objects near a given position.

Result:

```
SELECT p.objID, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z
,      n.distance
  FROM dbo.fGetNearbyObjEq(150.116321, 2.205830, 2.0) AS n
  JOIN PhotoObj AS pON n.objID = p.objID
 order by n.distance
```

This is correct, and even uses optimal distance function `fGetNearbyObjEq`

So, does it understanmd the schema? Can it compose SQL?

Not so quick

Further prompt: can you also find the 2mass magnitudes for these objects?

Result:

```
SELECT p.objID, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z
, tm.j_m, tm.h_m, tm.k_m
, n.distance
FROM dbo.fGetNearbyObjEq(150.116321, 2.205830, 2.0) AS n
JOIN PhotoObj AS p
 ON p.objID = n.objID
LEFT JOIN TwoMass AS tm
 ON p.objID = tm.objID
ORDER BY n.distance;
```

SDSS error message

Invalid column name 'j_m'. Invalid column name 'h_m'. Invalid column name 'k_m'.

Text-to-approximate-SQL, for SDSS

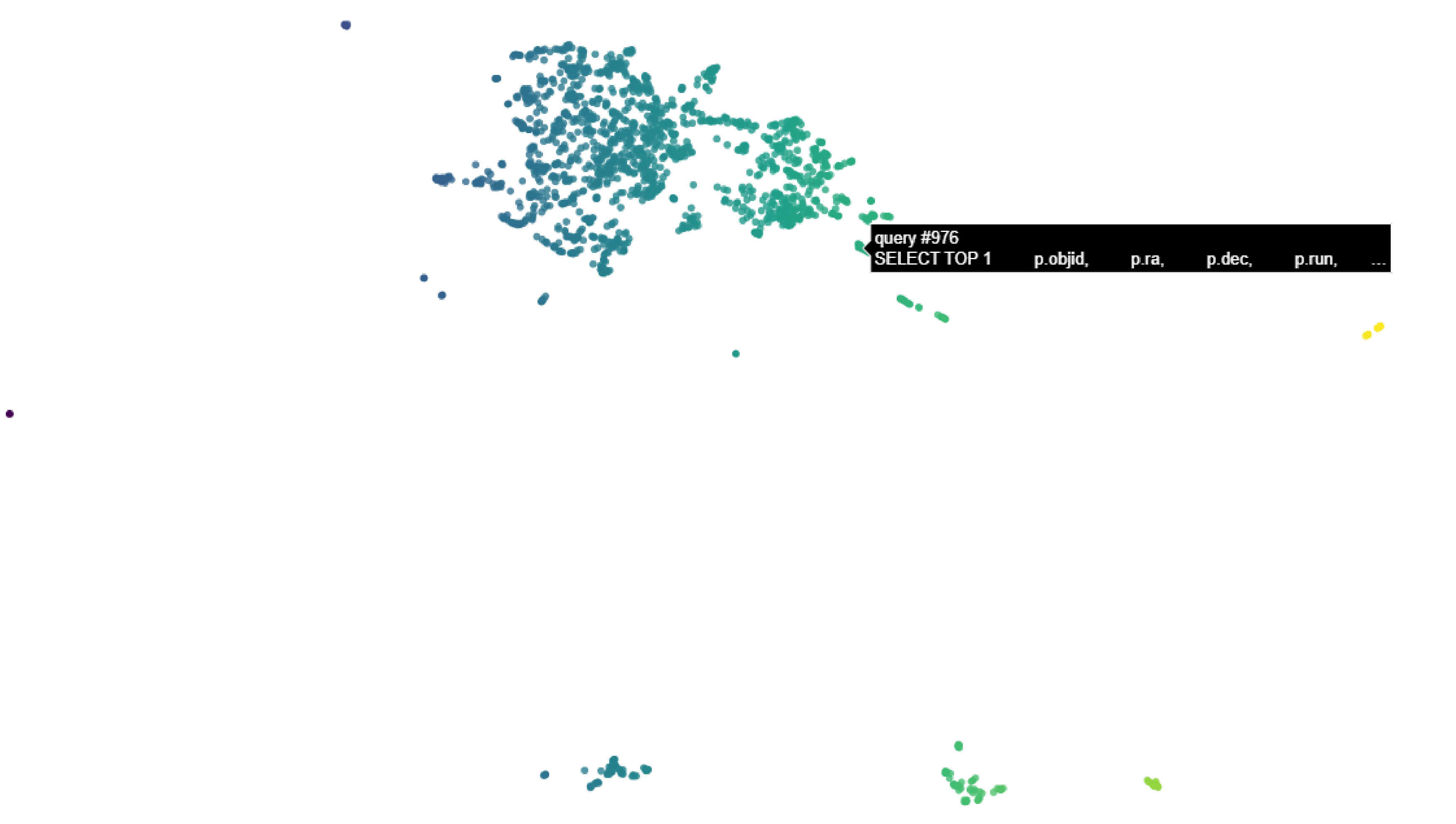
- Use corpus of SDSS SQL queries
  - 40M (CasJobs) or 1B (SkyServer) depending on how you count
  - No Text version
- Use Text *derived from* SQL to create embeddings
  - Using Gemma 4
  - Use: Find few nearest English descriptions and the corresponding SQL, show to user
  - Next: ask an LLM to create a proper "mean" query

OR:

- Create embeddings from SQL
  - Open source model, can be used on SciServer
    - Octen/Octen-Embedding-4B
    - On a100
  - store in vector DB
  - Seems that this is already good enough to find SQL from English language

Text-to-SQL: Clustering SQL queries

Query UMAP embedding — click a point to see its query



Query #976

```
    SELECT TOP 1
        p.objid,
        p.ra,
        p.dec,
        p.run,
        p.rerun,
        p.camcol,
        p.field,
        SQRT(POWER(p.ra - 1.00928006988209, 2) +
POWER(p.dec - -0.792000521668464, 2)) AS dist_deg
    FROM photoObj AS p
    WHERE
        p.ra BETWEEN 1.0090022921043122 AND
1.009557847659868
        AND p.dec BETWEEN -0.7922782994462418 AND
-0.7917227438906862
    ORDER BY dist_deg ASC
    ,SELECT TOP 1
        p.objid,
        p.ra,
        p.dec,
        p.run,
        p.rerun,
        p.camcol,
        p.field,
        SQRT(POWER(p.ra - 1.00928006988209, 2) +
POWER(p.dec - -0.792000521668464, 2)) AS dist_deg
    FROM photoObj AS p
    WHERE
        p.ra BETWEEN 1.0090022921043122 AND
1.009557847659868
        AND p.dec BETWEEN -0.7922782994462418 AND
-0.7917227438906862
    ORDER BY dist_deg ASC
```

Text-to-SQL continued

- What to do if we don’t have a large enough SQL corpus, say for new databases or 100 tables added to existing DB (DR19 -> DR20)
  - we have the fully documented schems
  - BUT are restricted to local LLM models
  - GenAI smart enough to translate?
- Has someone done this already?
  - After long deliberations Claude concludes there are no off the shelf solutions that do not require either a frontier model agentic system, or a set of ~100 SQL queries
- Will it ever be good enough to replace SQL
  - Is English precise enough?
- Is it necessary?
  - It is good to know SQL, like math for physics
  - Text-to-SQL as individualized Quick Start

AI with data on SciServer

- Create “AI-ready”, multi-modal datasets
- Use xmatch to determine tuples of sources from multiple catalogues.
- For many catalogues we have extra data: images, spectra
- Host and share results (for others) to use for training

AI ready (?), trusted (?) astronomy datasets

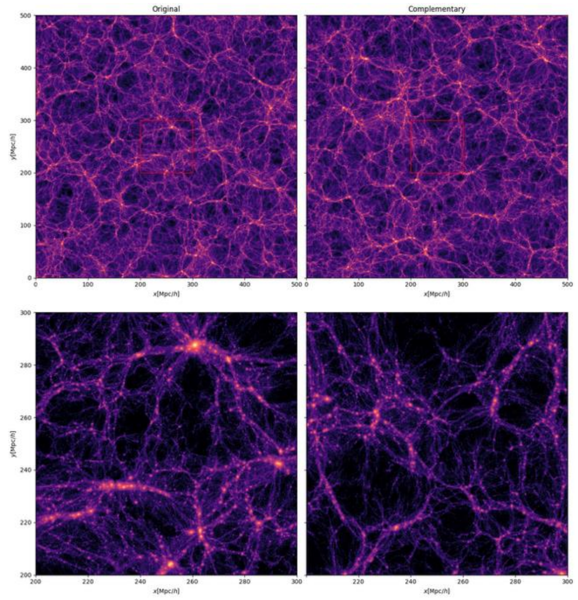
SDSS

- Sloan Digital Sky Survey (SDSS) Catalog Data (~150TB)
- SDSS Science Archive Server (SAS) Data (400TB)
- SDSS-IV eBOSS Spectra
- SDSS-IV MaNGA Integral Field Unit (IFU) Spectra
- SDSS Associated Data
- APOGEE FIRE Mock Catalogs
- SDSS-I/-II Data Archive Server

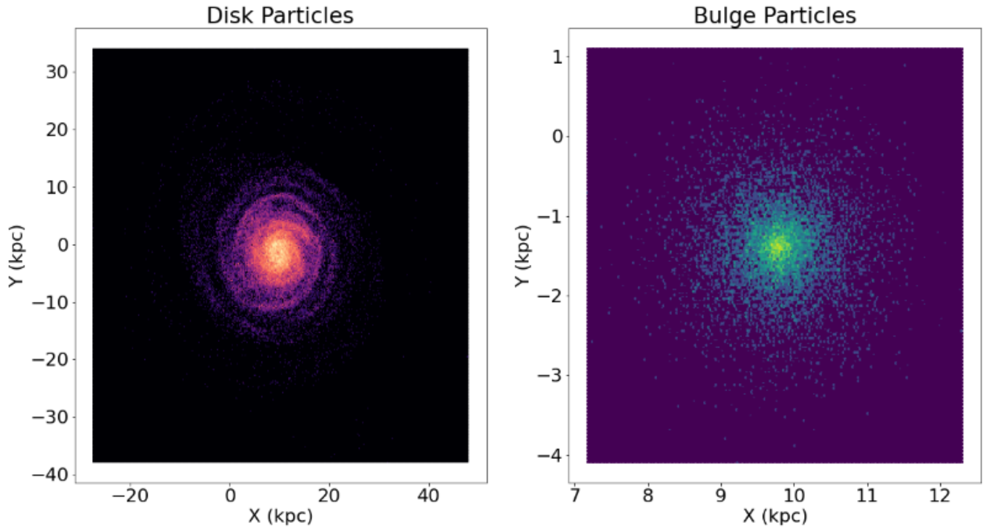
OTHER

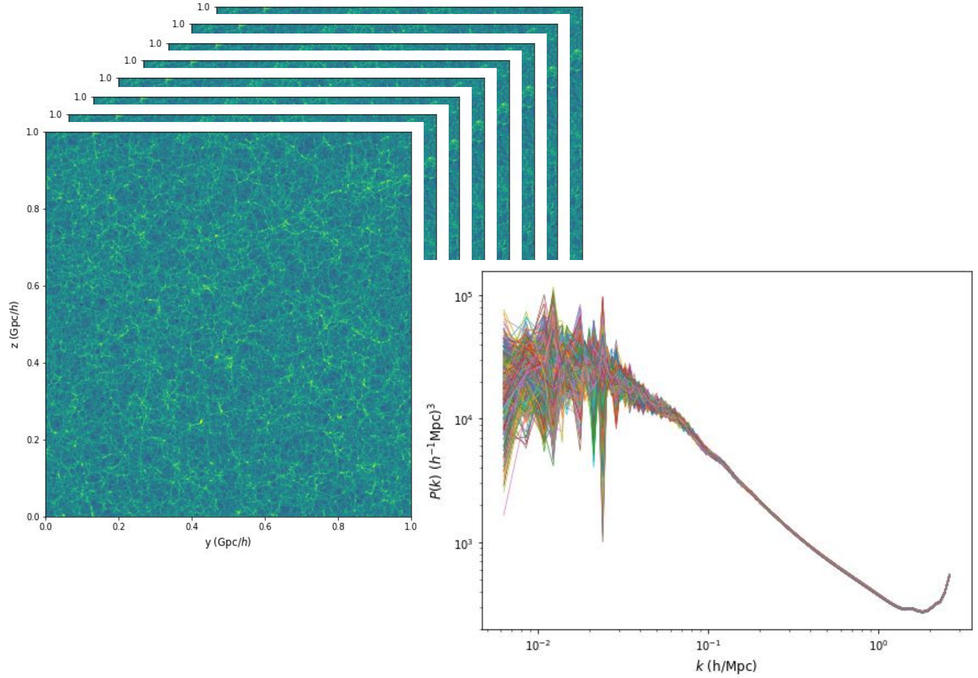
- XMATCH: Cross-matching astronomical datasets
- Faint Images of the Radio Sky at Twenty cm (FIRST)
- Two-Micron All-Sky Survey (2MASS)
- Gaia
- Sky Mapper Southern Survey (SMSS)
- Minor Planet Center orbital elements (MPCORB)
- Galactic Archaeology with HERMES (GALAH) DR4
- LAMOST DR8v2
- Two-Degree Field (2DF) Galaxy Redshift Survey
- Galaxy Evolution Explorer (GALEX)
- **High-Energy Astrophysics (HEASARC) (~200TB)**
- DESI DR1 spectra (~300TB)
- DES
- … more to come

Cosmological Simulations in the same platform



- Millennium: RDB catalogues + raw data (70TB)
  - Complementary Millennium Razc etal
- Indra: 468 1B particle simulations (~750TB)
- SMUDGE: galaxy + satellite merger simulations (145TB)





JOHNS HOPKINS UNIVERSITY | Data Science and AI Institute

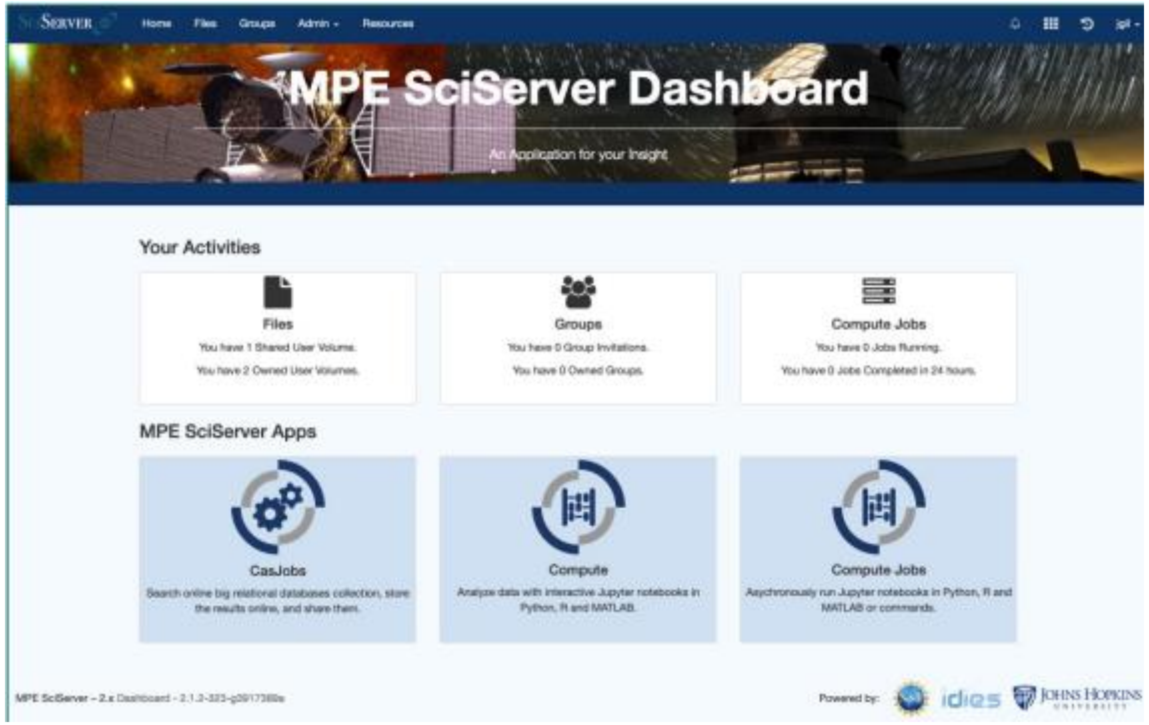
idies The Institute for Data-Intensive Engineering and Science

Bring us your …

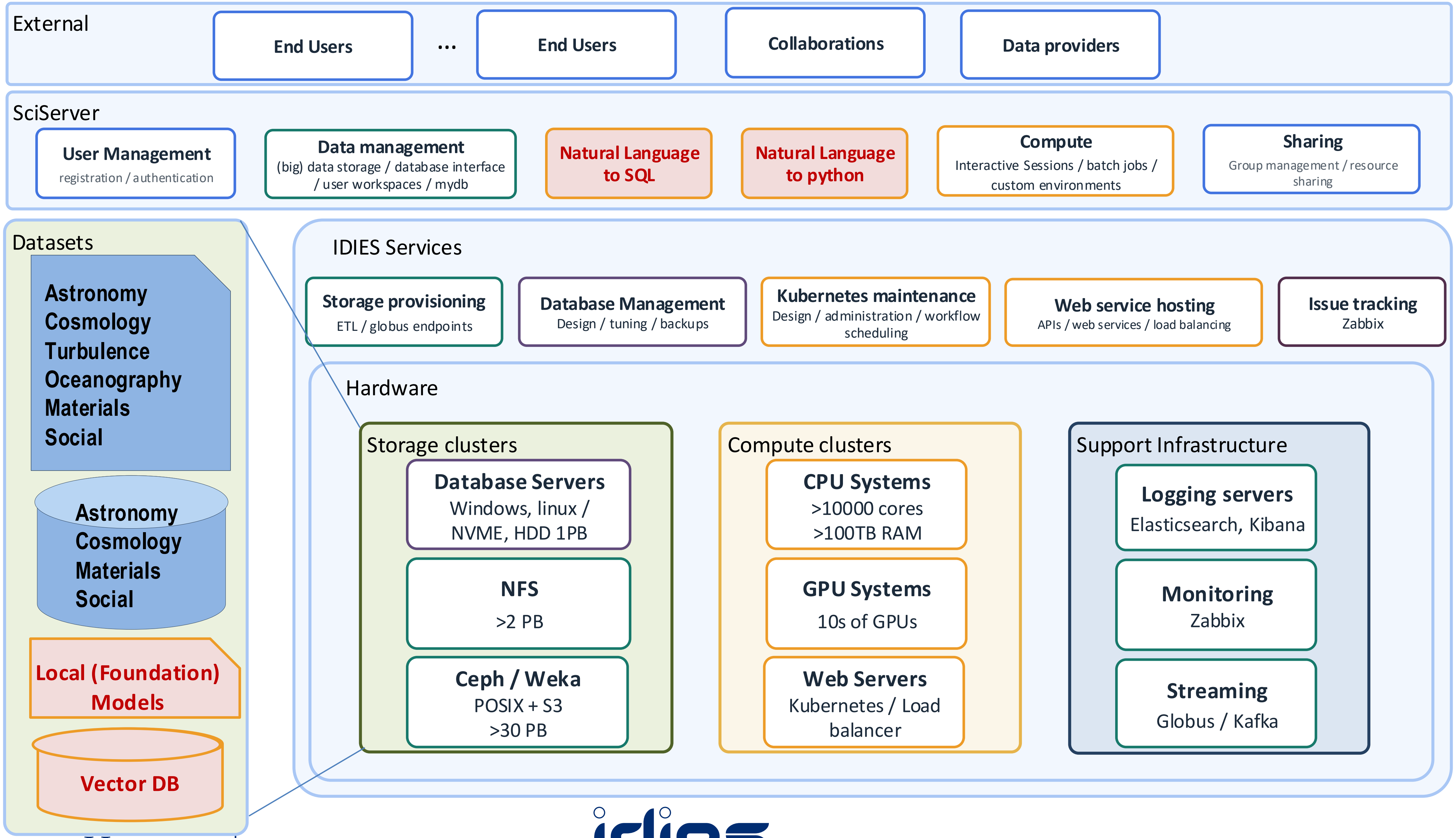
- Data
  - 1-5PB available
  - Free if <10TBs for dissemination of interesting datasets
  - If >>10TB, ~2k/100TB/yr with MoU etc
- Models
  - With code
- Solutions
  - Text-to-sql?
- Embedding vectors
- Anything else ??

Or deploy your own SciServer instance

- Open source: https://github.com/sciserver/opensciserver
  - Thanks to NSF grant OpenSciServer
- E.g. SciServer @ MPE
  - eROSITA
  - HETDEX

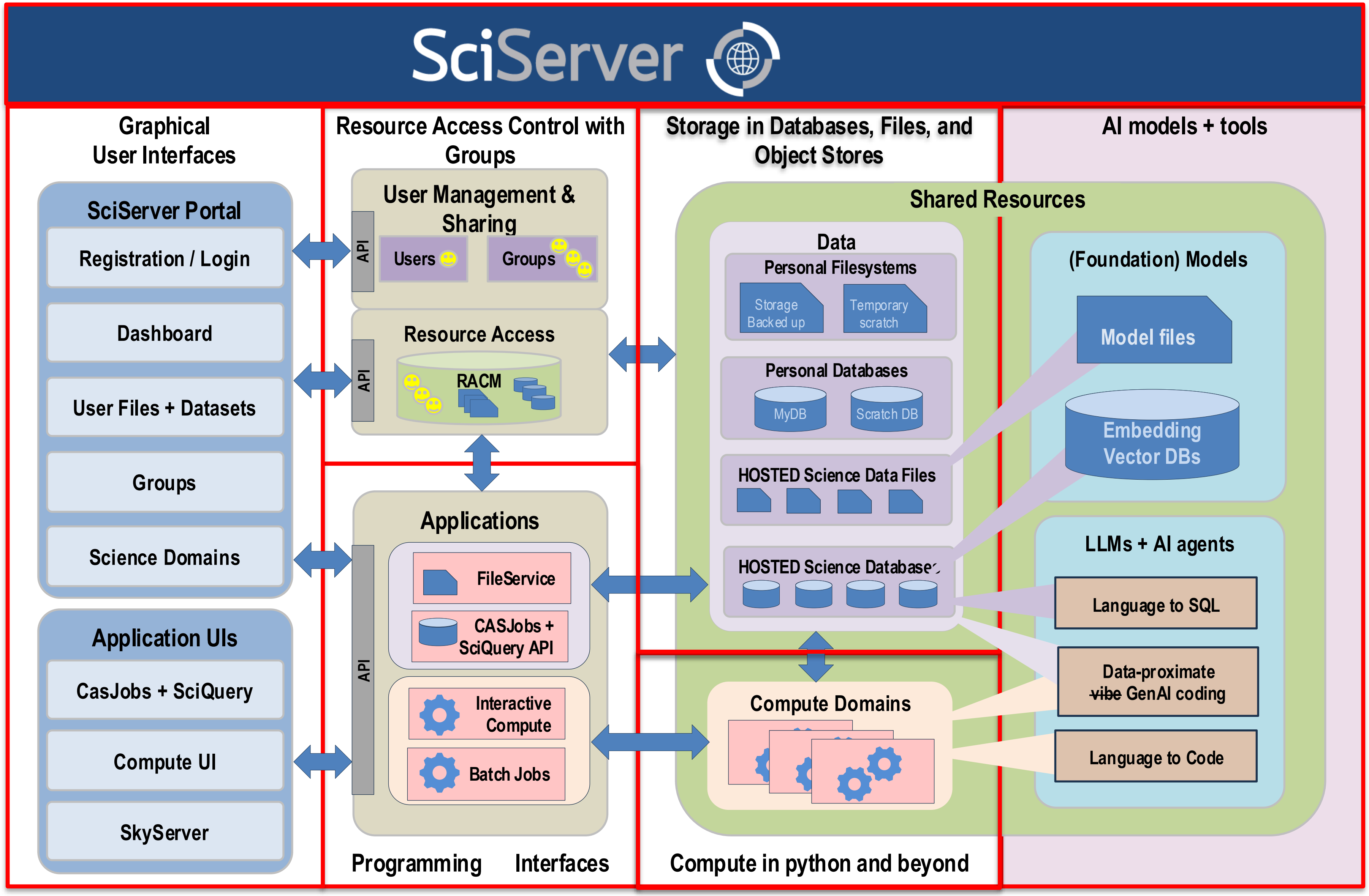


Thank you



JOHNS HOPKINS UNIVERSITY | Data Science and AI Institute

idies
The Institute for Data-Intensive Engineering and Science



Johns Hopkins University | Data Science and AI Institute

idies
The Institute for Data-Intensive Engineering and Science