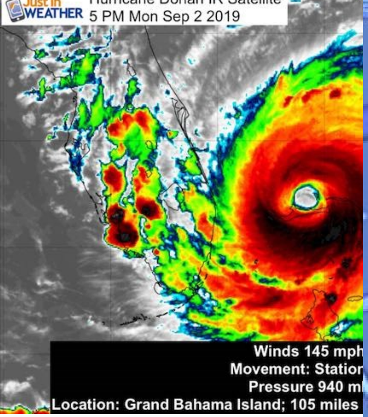
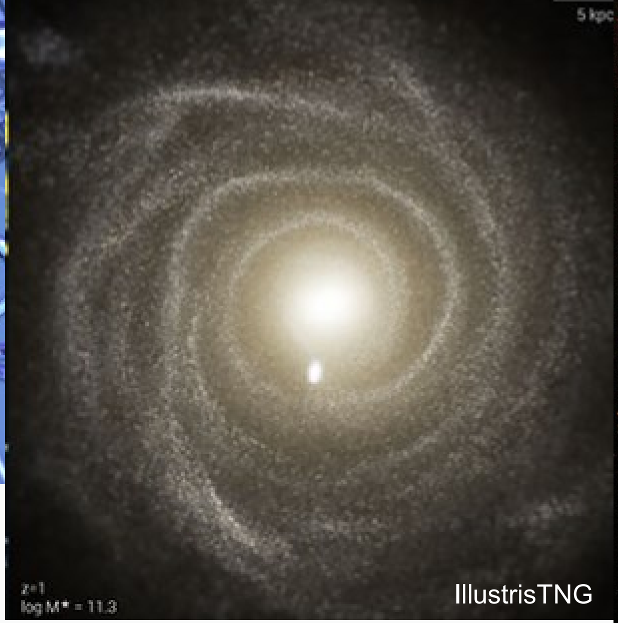


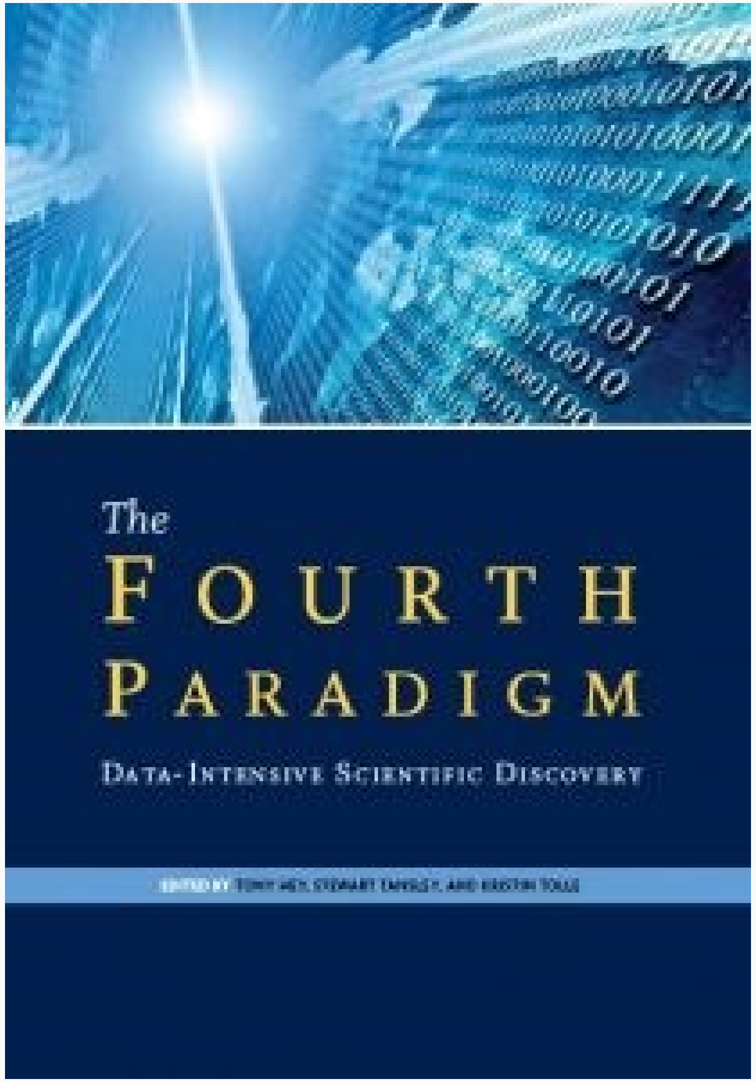



SciServer, a collaborative science platform with cosmological simulations

Gerard Lemson
Institute for Data Intensive Engineering and Science (IDIES)
The Johns Hopkins University

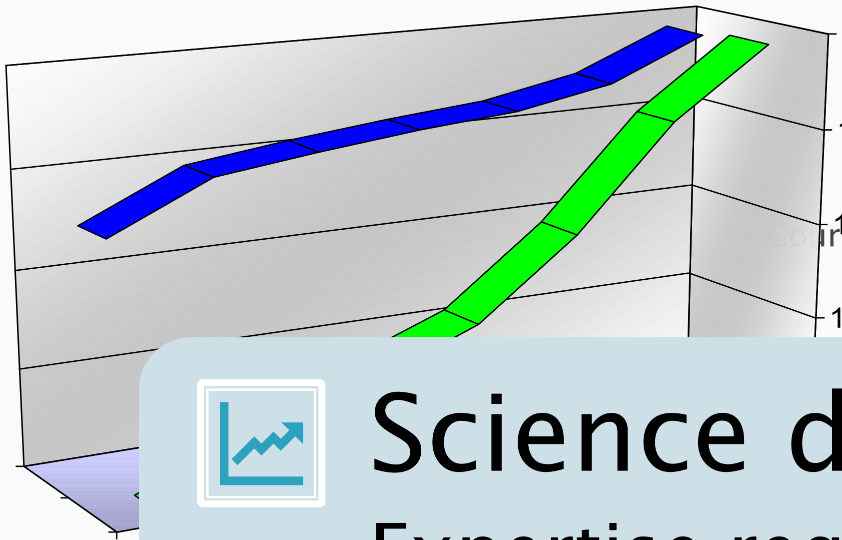
The 4th paradigm : data intensive scientific discovery

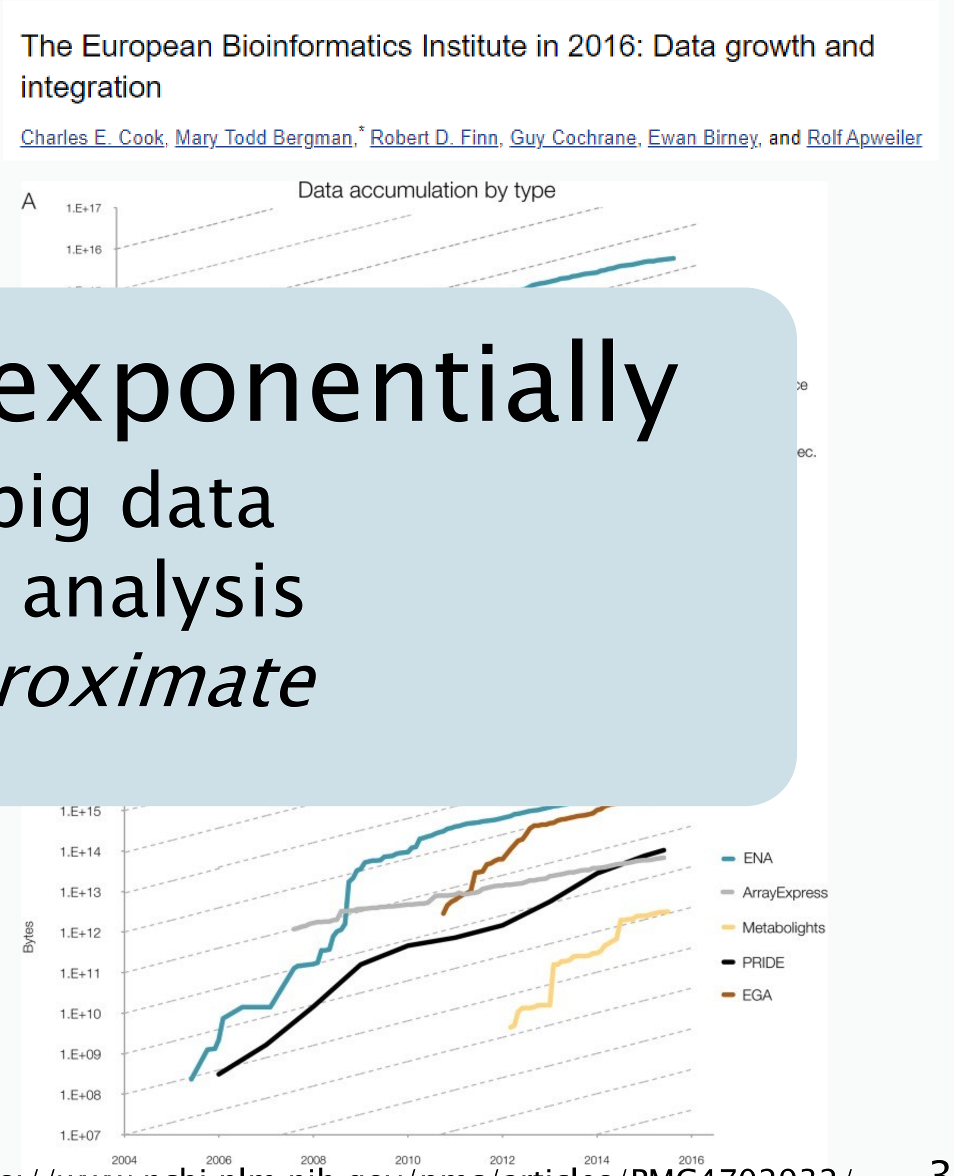
- Jim Gray



Increasingly scientific investigations require **combination of large amounts of data** from many different sources and ever more **sophisticated machine learning algorithms and tools** for their analysis.

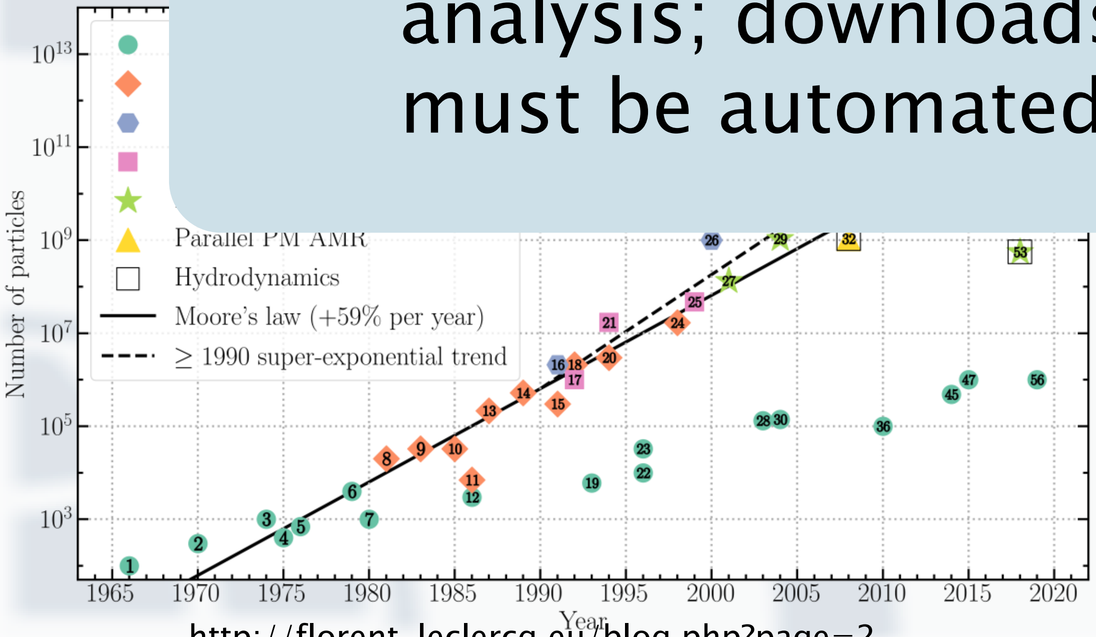
SciServer





Science data growing exponentially

Expertise required to optimize big data analysis; downloads unfeasible; analysis must be automated and *data-proximate*



http://florent-leclercq.eu/blog.php?page=2

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702932/

Planck 2018 results. IX. Constraints on primordial non-Gaussianity

Show affiliations | Hide authors

Planck Collaboration ; Akrami, Y. ; Arroja, F. ; Ashdown, M. ;
Aumont, J. ; Baccigalupi, C. ; Ballardini, M. ; Banday, A. J. ;
Barreiro, R. B. ; Bartolo, N. ; Basak, S. ; Benabed, K. ; Bernard, J. -P. ;
Bersanelli, M. ; Bielewicz, P. ; Bond, J. R. ; Borrill, J. ; Bouchet, F. R. ;
Bucher, M. ; Burigana, C. ; Butler, R. C. ; Calabrese, E. ; Cardoso, J. -F. ;
Casaponsa, B. ;
Crill, B. P. ; Cu
Delabrouille, J
Douspis, M. ;
Enßlin, T. A.
Fernandez-Co
Franceschi, E.
Génova-Santo
Gratton, S. ; G
Handley, W.
Jaffe, A. H.
Kiiveri, K. ; Ki
Lamarre, J. -M
Levrier, F. ; Le
López-Canieg
Maino, D. ; Ma
Martin, P. G.
McEwen, J. D.
Mennella, A. ; Migliaccio, M. ; Miville-Deschênes, M. -A. ; Molinari, D. ;
Moneti, A. ; Montier, L. ; Morgante, G. ; Moss, A. ; Münchmeyer, M. ;
Natoli, P. ; Oppizzi, F. ; Pagano, L. ; Paoletti, D. ; Partridge, B. ;
Patanchon, G. ; Perrotta, F. ; Pettorino, V. ; Piacentini, F. ; Polenta, G. ;
Puget, J. -L. ; Rachen, J. P. ; Racine, B. ; Reinecke, M. ; Remazeilles, M. ;
Renzi, A. ; Rocha, G. ; Rubiño-Martín, J. A. ; Ruiz-Granados, B. ;
Salvati, L. ; Savelainen, M. ; Scott, D. ; Shellard, E. P. S. ; Shiraishi, M. ;
Sirignano, C. ; Sirri, G. ; Smith, K. ; Spencer, L. D. ; Stanco, L. ; Sunyaev, R. ;
Suur-Uski, A. -S. ; Tauber, J. A. ; Tavagnacco, D. ; Tenti, M. ; Toffolatti, L. ;
Tomasi, M. ; Trombetti, T. ; Valiviita, J. ; Van Tent, B. ; Vielva, P. ;
Villa, F. ; Vittorio, N. ; Wandelt, B. D. ; Wehus, I. K. ; Zacchei, A. ;
Zonca, A.

The Sloan Digital Sky Survey: Technical Summary

Show affiliations | Hide authors

York, Donald G. ; Adelman, J. ; Anderson, John E., Jr. ; Anderson, Scott F. ;
Annis, James ; Bahcall, Neta A. ; Bakken, J. A. ; Barkhouser, Robert ;
Bastian, Steven ; Berman, Eileen ; Boroski, William N. ; Bracker, Steve ;
Briegel, Charlie ; Briggs, John W. ; Brinkmann, J. ; Brunner, Robert ;
Burles, Scott ; Carey, Larry ; Carr, Michael A. ; Castander, Francisco J. ;
Chen, Bing ; Colestock, Patrick L. ; Connolly, A. J. ; Crocker, J. H. ;
Okamura, Sadanori ; Ostriker, Jeremiah P. ; Owen, Russell ;
Pauls, A. George ; Peoples, John ; Peterson, R. L. ; Petravick, Donald ;
Pier, Jeffrey R. ; Pope, Adrian ; Pordes, Ruth ; Prosapio, Angela ;
Rechenmacher, Ron ; Quinn, Thomas R. ; Richards, Gordon T. ;
Richmond, Michael W. ; Rivetta, Claudio H. ; Rockosi, Constance M. ;
Ruthmansdorfer, Kurt ; Sandford, Dale ; Schlegel, David J. ;
Schneider, Donald P. ; Sekiguchi, Maki ; Sergey, Gary ; Shimasaku, Kazuhiro ;
Siegmund, Walter A. ; Smee, Stephen ; Smith, J. Allyn ; Snedden, S. ;
Stone, R. ; Stoughton, Chris ; Strauss, Michael A. ; Stubbs, Christopher ;
SubbaRao, Mark ; Szalay, Alexander S. ; Szapudi, Istvan ;
Szokoly, Gyula P. ; Thakar, Anirudda R. ; Tremonti, Christy ;
Tucker, Douglas L. ; Uomoto, Alan ; Vanden Berk, Dan ;
Vogeley, Michael S. ; Waddell, Patrick ; Wang, Shu-i. ; Watanabe, Masaru ;
Weinberg, David H. ; Yanny, Brian ; Yasuda, Naoki ; SDSS Collaboration

Full length article

The illustris simulation: Public data release ☆

D. Nelson[a], A. Pillepich[a], S. Genel[b, a, 1], M. Vogelsberger[c], V. Springel[d, e], P. Torrey[c, g], V. Rodriguez-Gomez[a], D. Sijacki[f], G.F. Snyder[h], B. Griffen[c], F. Marinacci[c], L. Blecha[j, 2], L. Sales[i], D. Xu[d], L. Hernquist[a]

g TAPIR, Mailcode 350-17, California Institute of Technology, Pasadena, CA 91125, USA

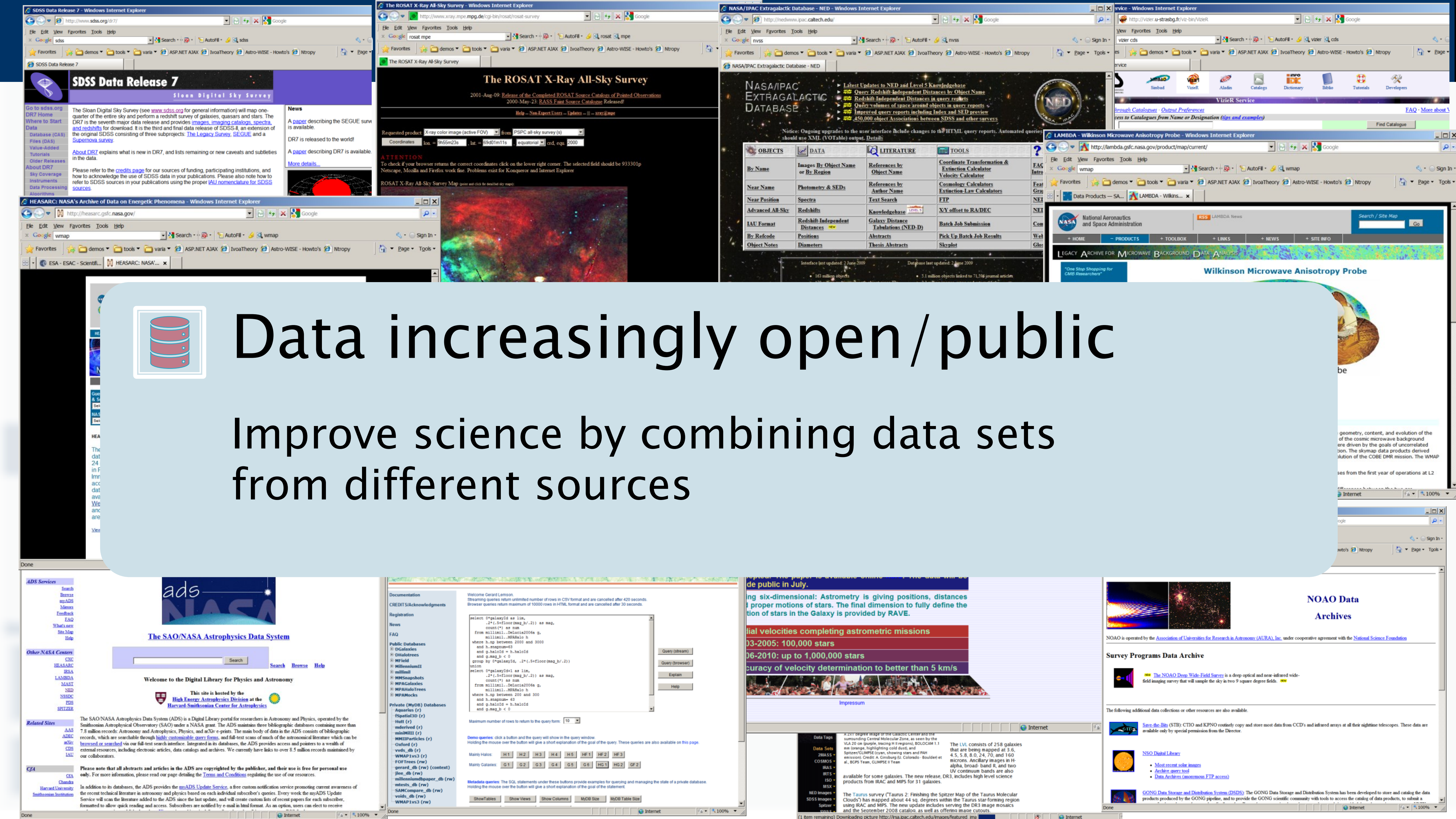
h Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218, USA

i Department of Physics and Astronomy, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA

j University of Maryland, College Park, Department of Astronomy and Joint Space Science Institute, USA

Science increasingly collaborative

Data and analysis sharing often ad hoc



Data increasingly open/public

Improve science by combining data sets from different sources

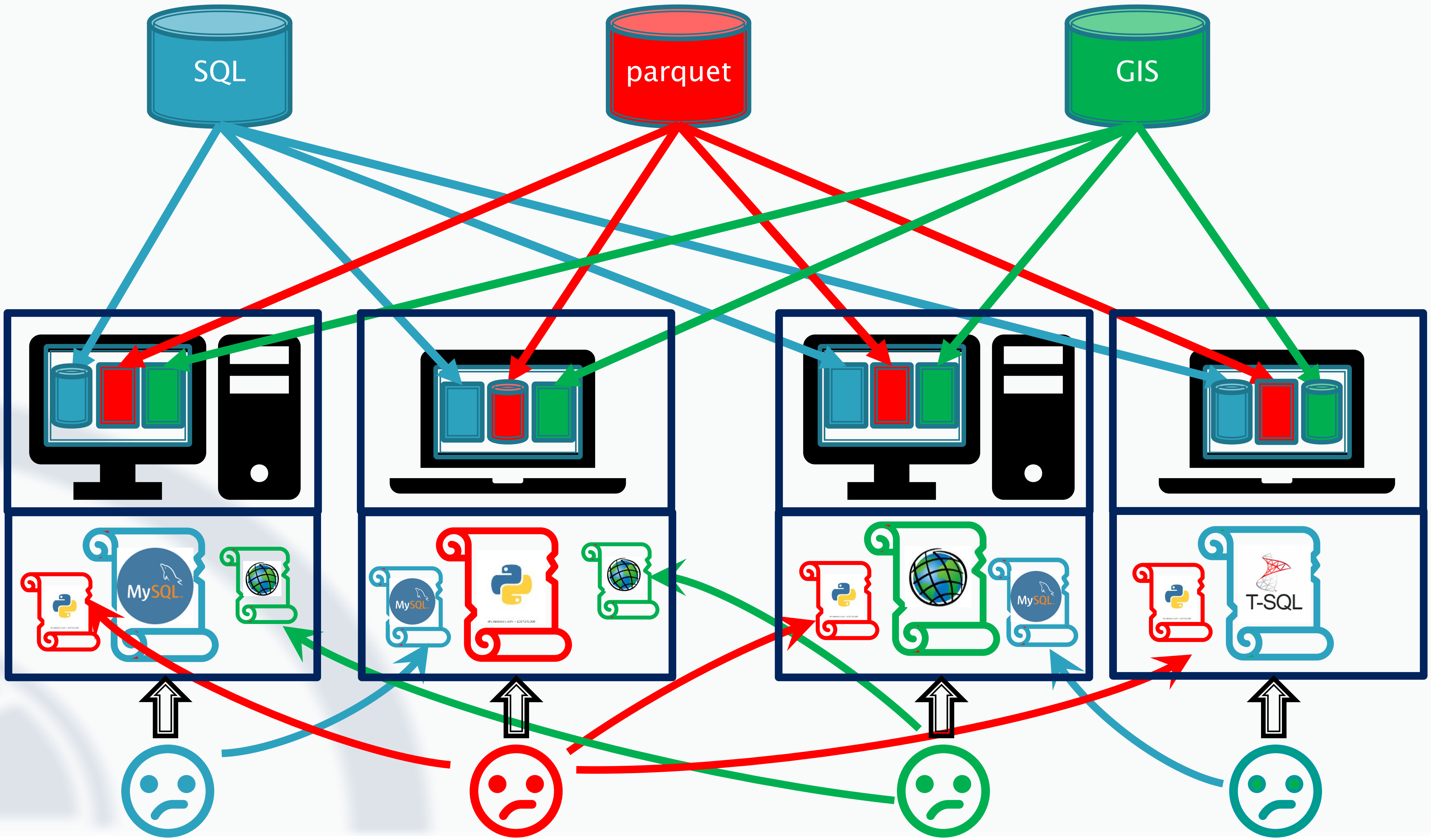
Data sources/formats heterogeneous

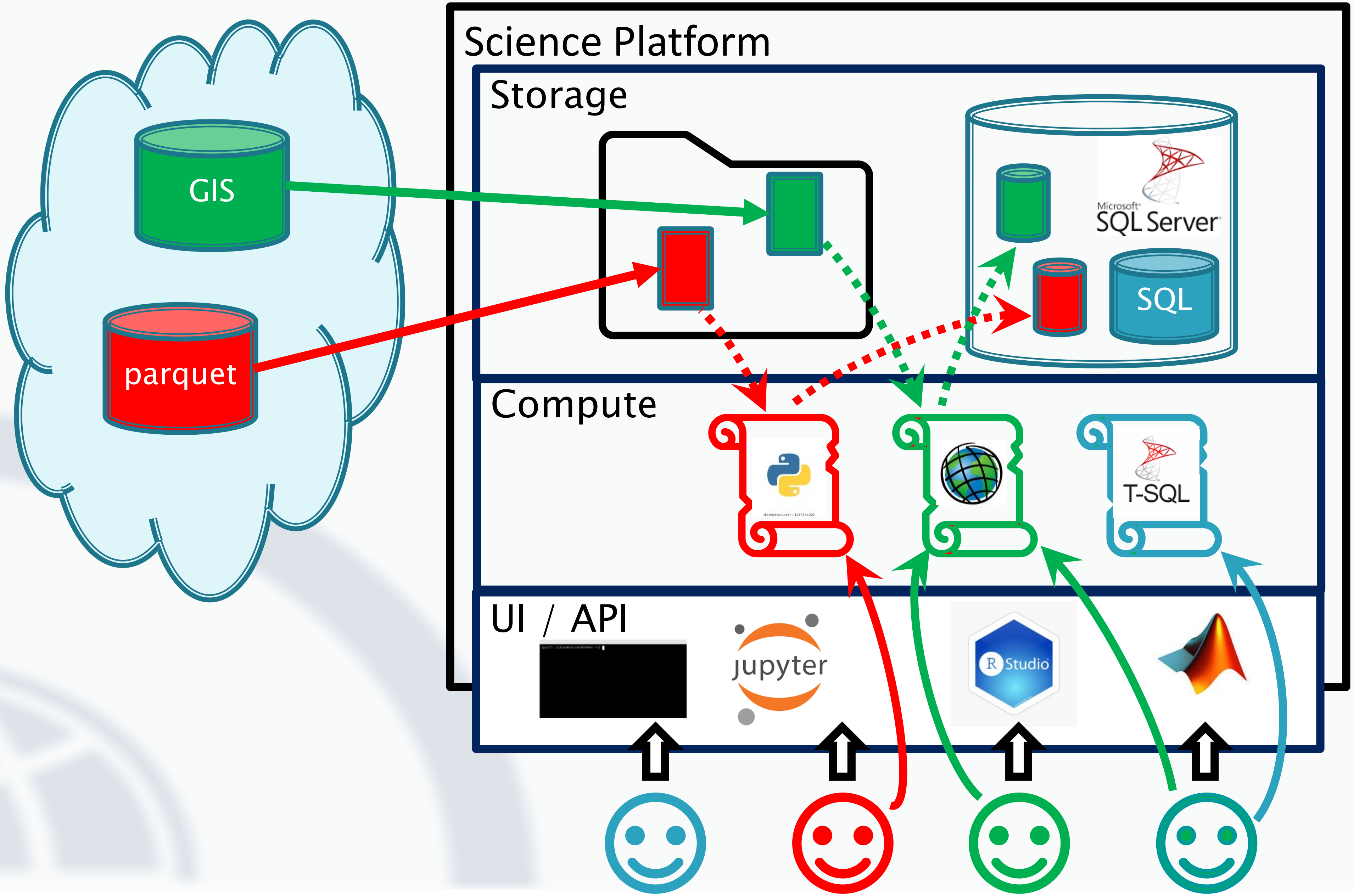
Diverse set of skills and knowledge required, both technical and domain knowledge

Question:

- How can we improve the support for **collaborative science** projects producing and requiring **large, heterogeneous data** sets with **geographically distributed** partners with **varying expertise**?

SciServer




Science Platform
Storage
GIS
parquet
Microsoft SQL Server
SQL
Compute
T-SQL
UI / API
jupyter
R Studio

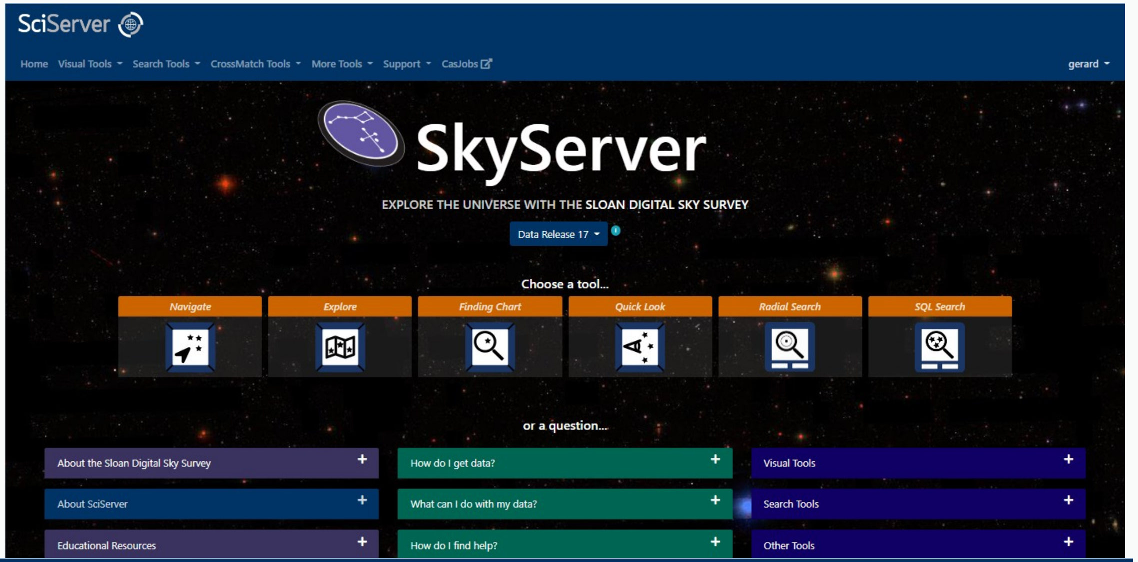
At JHU/IDIES we started with SkyServer

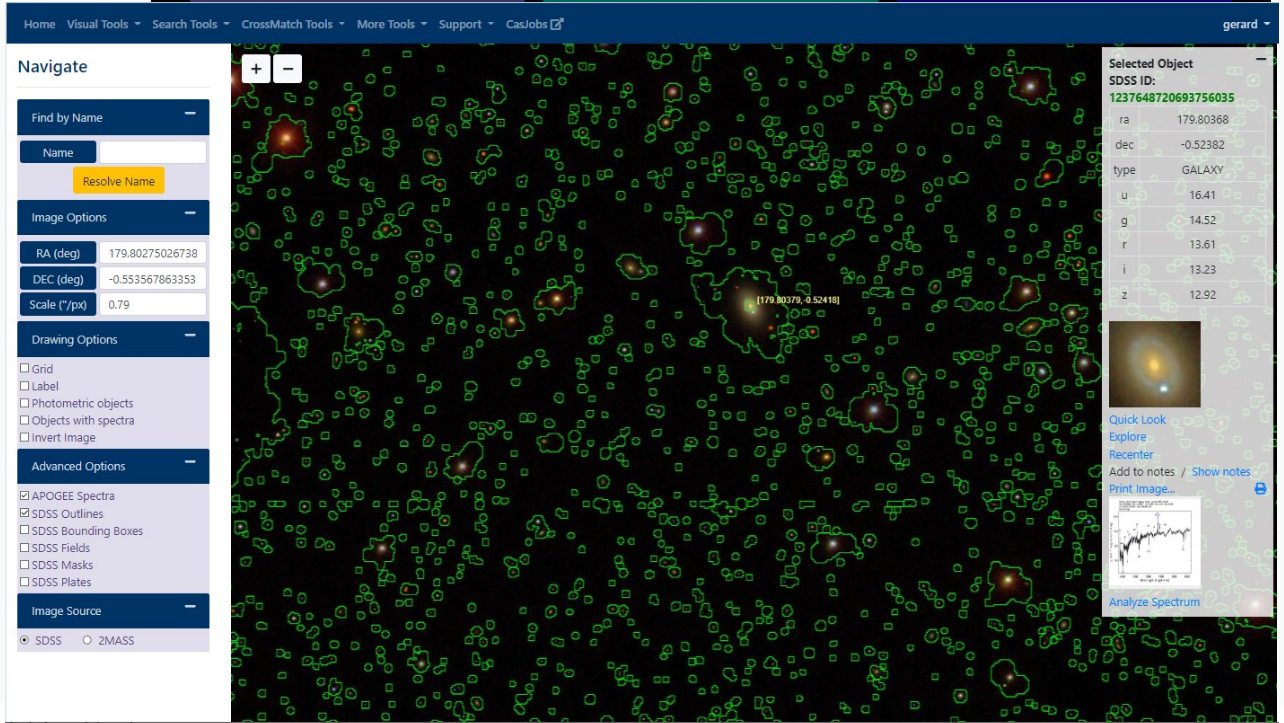
Disseminating data from Sloan Digital Sky Survey

<u>Goal</u>: **instant access** to rich content

<u>Idea</u>: **bring the analysis to the data**

Interactive access at the core



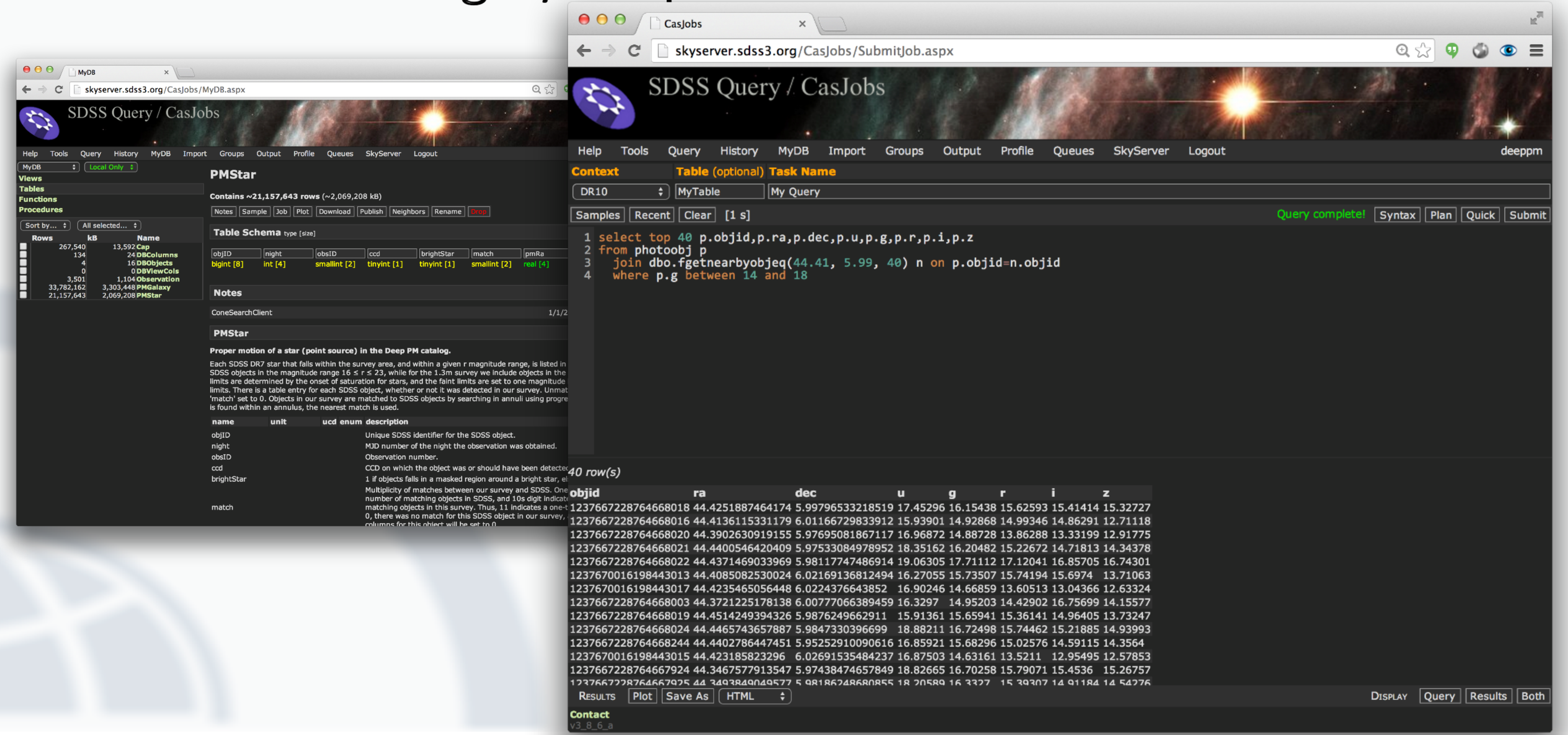


SLOAN DIGITAL SKY SURVEY: EARLY DATA RELEASE

Stoughton etal, 2002

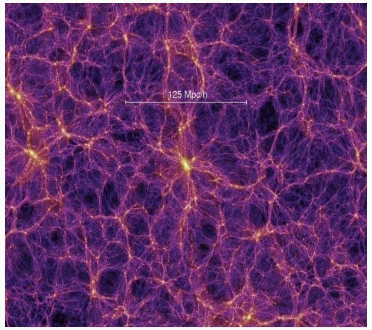
https://ui.adsabs.harvard.edu/abs/2002AJ....123..485S

But need for longer/deeper searches…



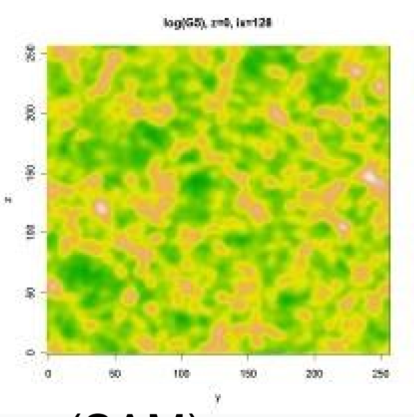
At MPA: Virgo-Millennium Database

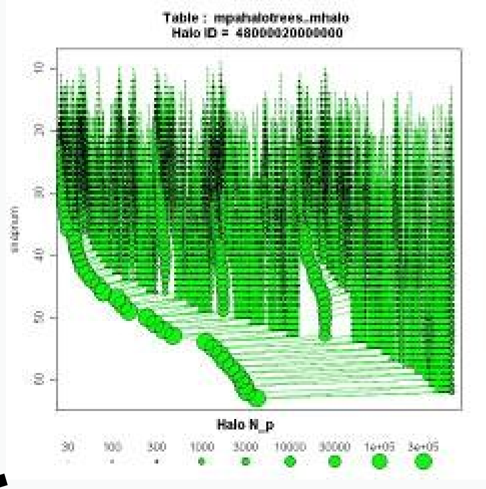
Raw data:
Particles



FOF groups
and Subhalos

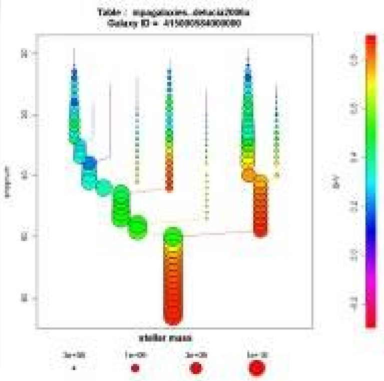
Density fields

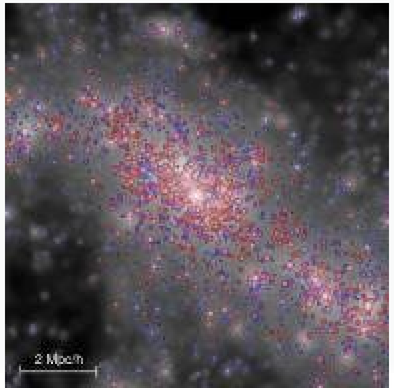


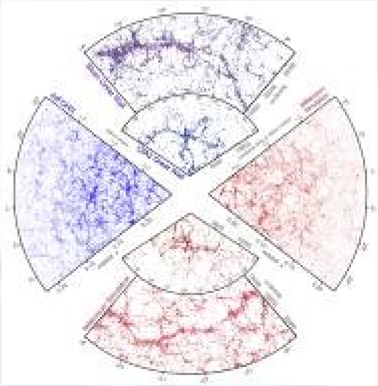


Subhalo merger trees

Synthetic galaxies (SAM)

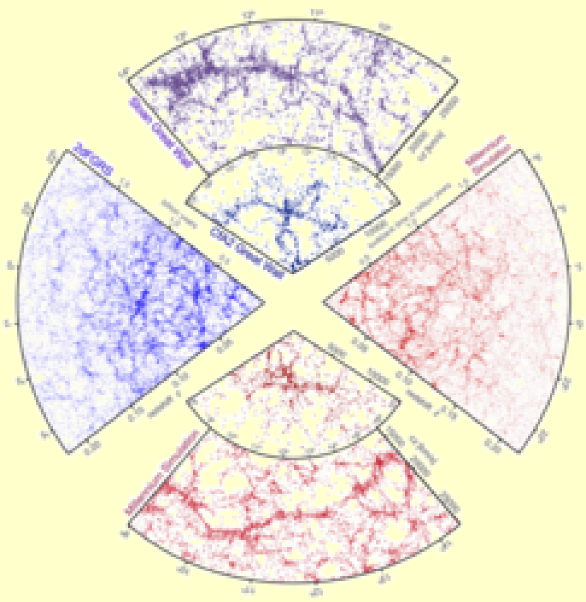






Mock catalogues

http://gavo.mpa-garching.mpg.de/Millennium/



- Database Access
- Visual Material
- Related Links
- Publications

Halo and Galaxy Formation Histories from the Millennium Simulation

Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the ΛCDM cosmogony

Gerard Lemson & the Virgo Consortium

astro-ph/0608019
full description of release (PDF)
database mirror site at ICC, Durham University

INTRODUCTION

When published in 2005, the Millennium Run was the largest ever simulation of the formation of structure within the ΛCDM cosmology. It uses 10^{10} particles to follow the dark matter distribution in a cubic region $500h^{-1}$Mpc on a side, and has a spatial resolution of $5h^{-1}$kpc. Application of simplified modelling techniques to the stored output of this calculation allows the formation and evolution of the

https://wwwmpa.mpa-garching.mpg.de/millennium/

Virgo - Millennium Database

- Documentation
- CREDITS/Acknowledgments
- Registration
- News
- FAQ

Public Databases
- Ayromlou2021a
- DGalaxies
- DHaloTrees
- Guo2010a
- Guo2013a
- Henriques2012a
  - Tables
    - wmap1.BC03_0ij
    - wmap1.BC03_AllSky_0ij
    - wmap1.M05_0ij
    - wmap1.M05_AllSky_0ij
    - wmap1_rest.BC03_0ij
    - wmap1_rest.M05_0ij
- Henriques2015a
  - Tables
    - cones.AllSky_M05_001
    - cones.AllSky_M05_002
    - cones.MRscPlanck1_BC03_0ij
    - cones.MRscPlanck1_M05_0ij
    - MRIIscPlanck1
    - MRscPlanck1
    - SFH_Times_MR
    - SFH_Times_MRII
- Henriques2020a
  - Tables
    - MRscPlanck1
    - MRscPlanck1_Rings
- MField
- MillenniumII
- millimil
- miniMilII
- MMSnapshots
- MPAGalaxies
- MPAHaloTrees
  - Tables
    - MHalo
    - MR
    - MR7
    - MRII
    - MRIIscPlanck1
    - MRIIscWMAP7
    - MRscPlanck1
    - MRscWMAP7
- MPAMocks
- Snapshots

Private (MyDB) Databases

Welcome Gerard Lemson.
Streaming queries return unlimited number of rows in CSV format and are cancelled after 420 seconds.
Browser queries return maximum of 1000 rows in HTML format and are cancelled after 30 seconds.

There is a partial mirror of this database in Durham at http://galaxy-catalogue.dur.ac.uk:8080/Millennium/.
The Durham database does not contain all the latest L-Galaxies models but does contain more recent GALFORM models.

```
select PROG.*
  from millimil..DeLucia2006a PROG,
       millimil..DeLucia2006a DES
 where DES.galaxyId = 1
   and PROG.galaxyId between DES.galaxyId and DES.lastprogenitorId
```

Query (stream) | Query (browser) | Explain | Help

Maximum number of rows to return to the query form: 10

Previous queries:
List of all queries executed sofar in this session. Selecting a query will make it appear in the query window.
The link will show all of them in a separate window. Refreshing that window will load the latest queries again.

select PROG.* from millimil..DeLucia2006a PROG, millimil..DeLucia20 — Show all previous queries

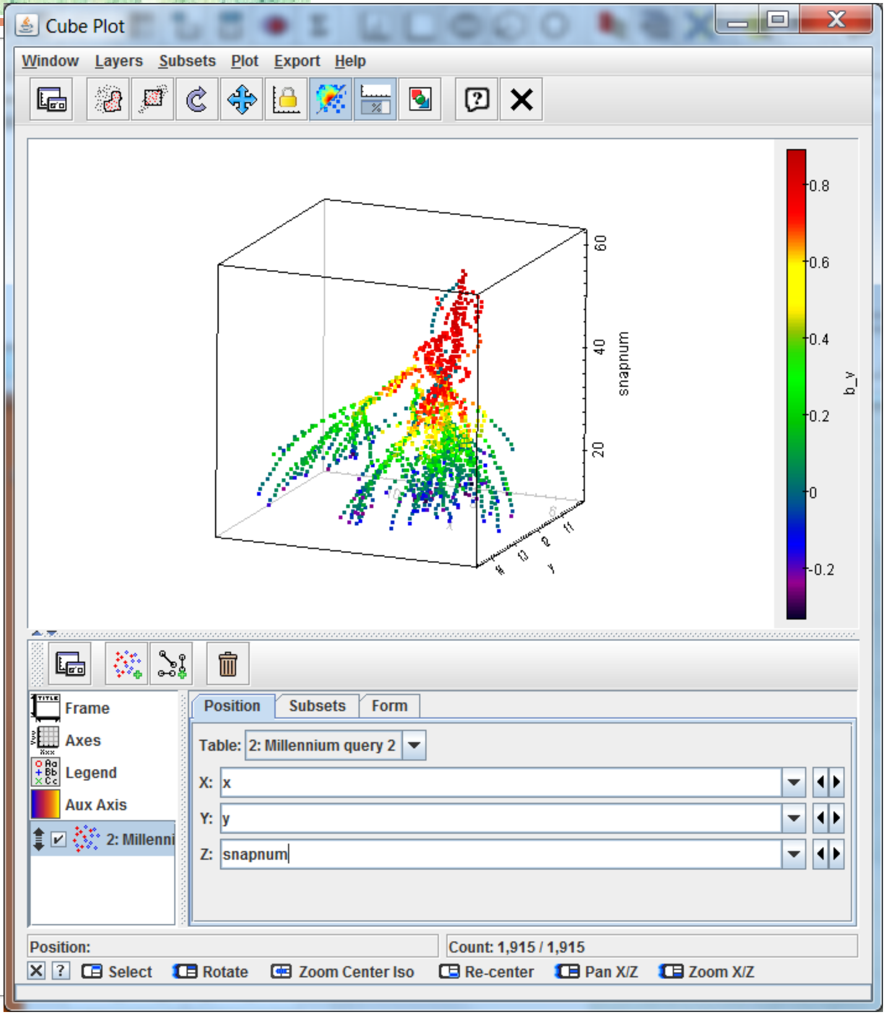
Demo queries: click a button and the query will show in the query window.
Holding the mouse over the button will give a short explanation of the goal of the query. These queries are described in some more detail on this page.

Mainly Halos: H 1 | H 2 | H 3 | H 4 | H 5 | HF 1 | HF 2 | HF 3

Mainly Galaxies: G 1 | G 2 | G 3 | G 4 | G 5 | G 6 | HG 1 | HG 2 | GF 2

Metadata queries: The SQL statements under these buttons provide examples for querying and managing the state of a private database.
Holding the mouse over the button will give a short explanation of the goal of the statement.

Show Tables | Show Views | Show Columns | Show Indexes | MyDB Size | MyDB Table Size
Create View | Drop Table | Create Index

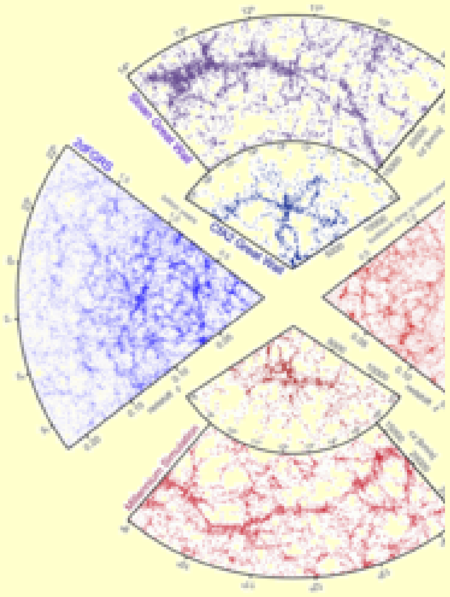


http://gavo.mpa-garching.mpg.de/Millennium/

Query time (in millisec) = 181
Number of rows retrieved from database = 1000

Reformat | CSV | Plot (VOPlot)

galaxyID	lastProgenitorId	descendantId	haloID	subHaloID	fofID	treeId	firstProgenitorId	nextProgenitorId	type	snapnum	redshift	centralMvir	phkey	x	y	z
1	1915	0	1	62000000000000	62000000000000	0	2	-1	0	62	0.019932542	3552.877	673	6.587909	13.099106	25
2	1869	1	2	61000000000000	61000000000000	0	3	1870	0	61	0.041403063	3494.8687	673	6.597178	13.111782	25
3	1400	2	3	60000000000000	60000000000000	0	4	1401	0	60	0.064493395	3318.8645	673	6.615912	13.121013	25
4	1337	3	4	59000000000000	59000000000000	0	5	1338	0	59	0.08928783	3285.643	673	6.6276503	13.1303835	25
5	1337	4	5	58000000000000	58000000000000	0	6	-1	0	58	0.11588337	3249.5815	673	6.6414022	13.1400175	25



- Database Acces
- Visual Materia
- Related Links
- Publications

distribution through galaxy-galaxy lensing in the Hyper Suprime-Cam survey
Authors: Wenting Wang, Xiangchong Li, Jingjing Shi, Jiaxin Han, Naoki Yasuda, Yipeng Jing, Surhud More, Masahiro Takada, Hironao Miyatake, Atsushi J. Nishizawa
Comments: submitted to ApJ - comments welcome - data available upon request

1088. astro-ph/2104.07664 [abs, ps, pdf, other]:
Title: Optimizing high redshift galaxy surveys for environmental information
Authors: Tobias J. Looser, Simon J. Lilly, Larry P. T. Sin, Bruno M. B. Henriques, Roberto Maiolino, Michele Cirasuolo
Comments: 29 pages, 40 figures. Accepted for publication in MNRAS

1089. astro-ph/2104.08295 [abs, ps, pdf, other]:
Title: The evolution of the mass-metallicity relations from the VANDELS survey and the GAEA Semi-Analytic model
Authors: Fabio Fontanot (1,2), Antonello Calabro (3), Margherita Talia (4,5), Filippo Mannucci (6), Marco Castellano (3), Giovanni Cresci (6), Gabriella De Lucia (1), Anna Gallazzi (6), Michaela Hirschmann (7), Laura Pentericci (3), Lizhi Xie (8), Ricardo Amorin (10,11), Micol Bolzonella (5), Angela Bongiorno (3), Olga Cucciati (5), Fergus Cullen (12), Johan P. U. Fynbo (13), Nimish Hathi (14), Pascale Hibon (15), Ross J. McLure (12), Lucia Pozzetti (5) ((1) INAF - Astronomical Observatory of Trieste, Italy (2) IFPU - Institute for Fundamental Physics of the Universe, Trieste, Italy (3) INAF - Astronomical Observatory of Rome, Italy (4) Dipartimento di Fisica e Astronomia, UniversitÃ¡ di Bologna, Italy (5) INAF - Astronomical Observatory of Bologna, Italy (6) INAF - Astrophysical Observatory of Arcetri, Firenze, Italy (7) DARK, Niels Bohr Institute, University of Copenhagen, Denmark (8) Tianjin Astrophysics Center, Tianjin Normal University, China (10) Instituto de InvestigaciÃ³n Multidisciplinar en Ciencia y TecnologÃa, Universidad de La Serena, Chile (11) Departamento de FÃsica y AstronomÃa, Universidad de La Serena, Chile (12) SUPA Scottish Universities Physics Alliance, Institute for Astronomy, University of Edinburgh, Royal Observatory (13) The Cosmic Dawn Center, Niels Bohr Institute, University of Copenhagen, Denmark (14) Space Telescope Science Institute, Baltimore, USA (15) ESO-Chile, Santiago, Chile)
Comments: 12 pages, 8 figures, MNRAS submitted

1090. astro-ph/2105.09126 [abs, ps, pdf, other]:
Title: Dynamical analysis of clusters of galaxies from cosmological simulations
Authors: Tania Aguirre Tagliaferro, Andrea Biviano, Gabriella De Lucia, Emiliano Munari, Diego Garcia Lambas
Comments: 14 pages, 17 figures

the Millennium

base for studying the
ogony

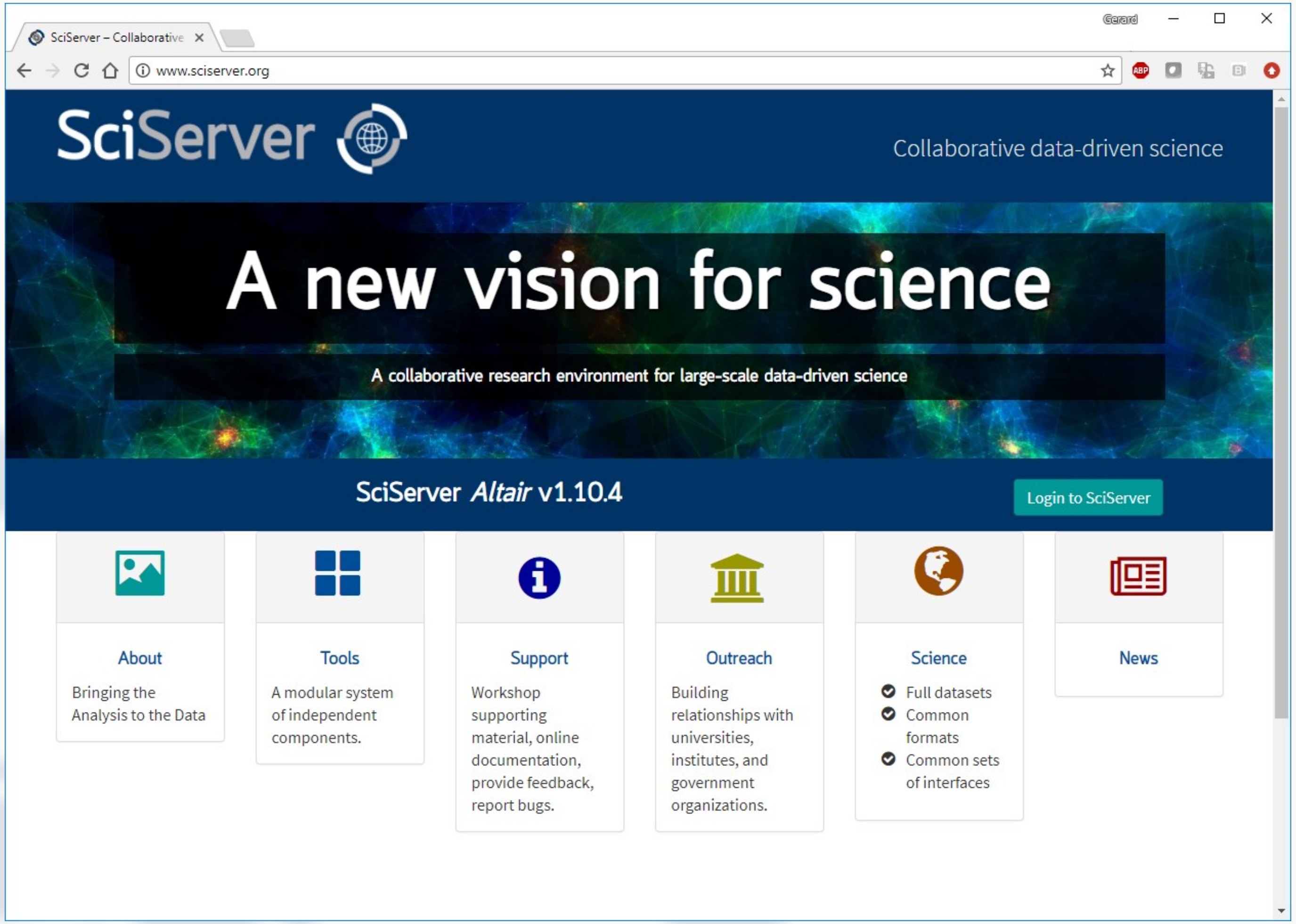
ty

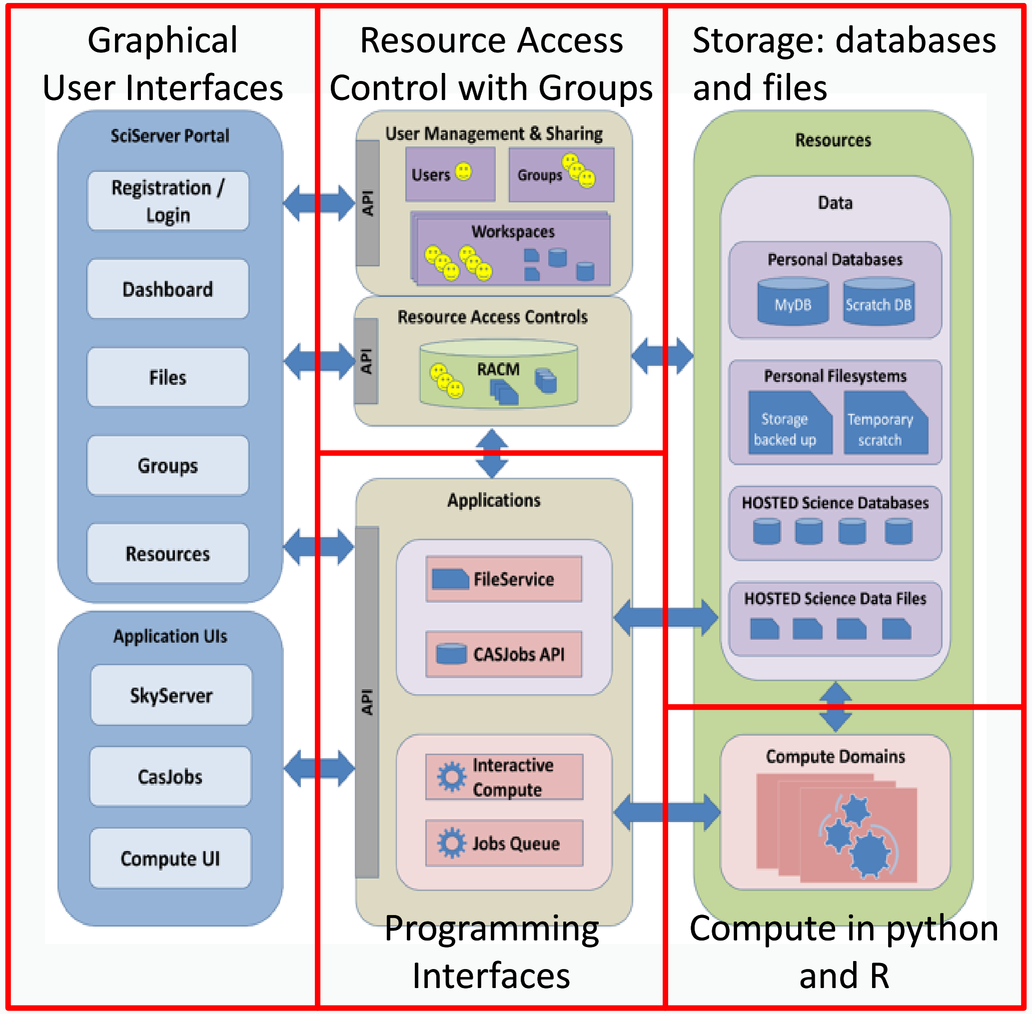
mulation of the formation of
dark matter distribution in a
kpc. Application of simplified
rmation and evolution of the

Users want more than “just” SQL:

- Want full analysis, visualization near the data
  - Run Python, R, C++, etc
- Want access to data that does not fit in a relational database
  - Images, spectra, data cubes, custom data sets
- Want to upload own data and combine
  - Need workspace close to data, databases and file system
- Want to share work with collaborators
  - Data, Scripts, Results
- Support for data and libraries from different disciplines
  - No single data model, ontologies

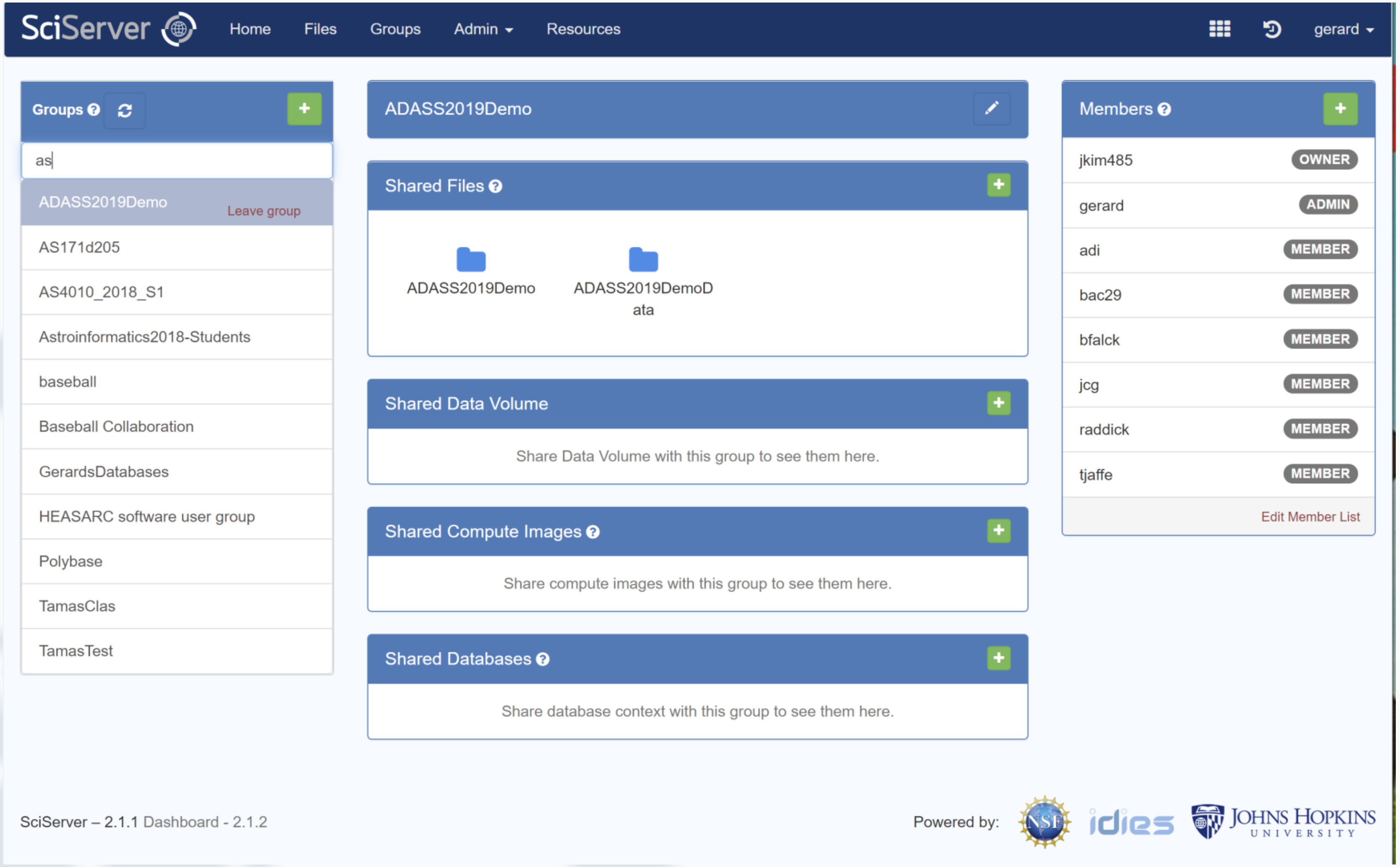
Collaborative data-driven science

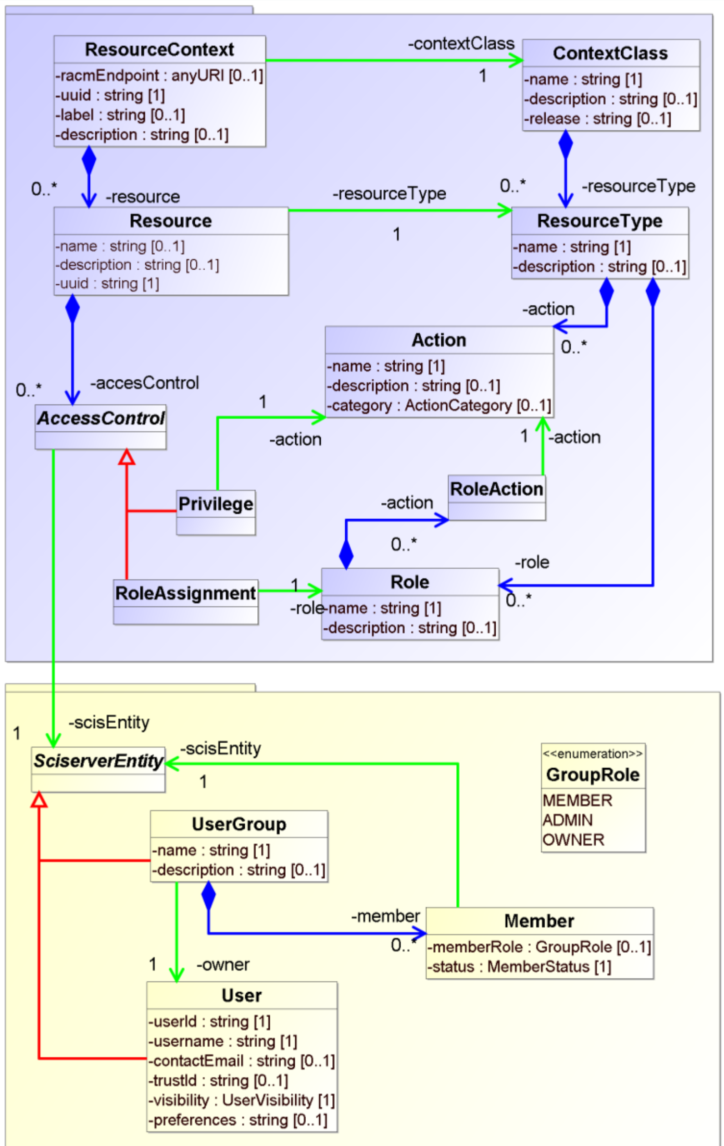




- **SciServer** is a system allowing Science Researchers across multiple domains to host and share their datasets, and provide query and analysis tools for collaborative research.
- **Core Services:**
  - Science Data Hosting (Files and Databases)
  - Query of hosted databases
  - Data Integration across hosted data sources
  - Computational analysis on hosted data
  - Collaboration and Sharing
  - Personal Storage (Files and Database)
  - API Integration

Resource sharing





SciServer Compute:
Jupyter (and more) in docker containers

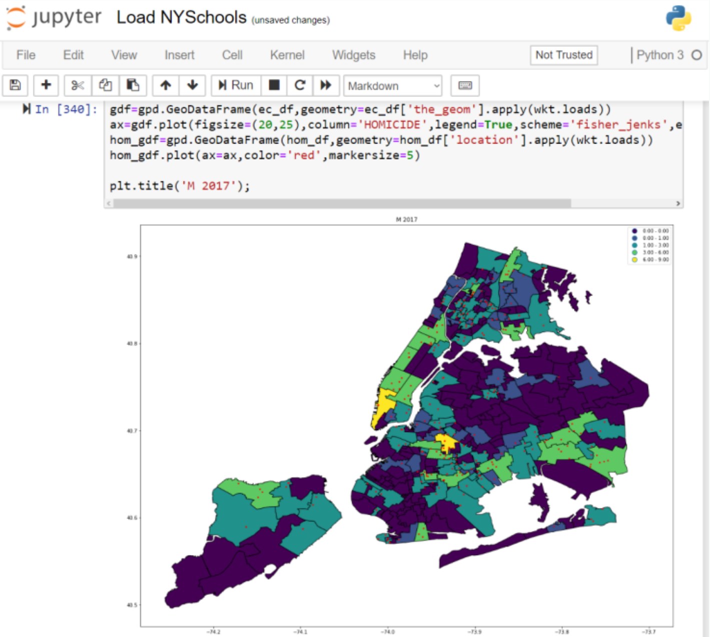
SciServer Compute | Interactive Notebooks | Jobs | gerard

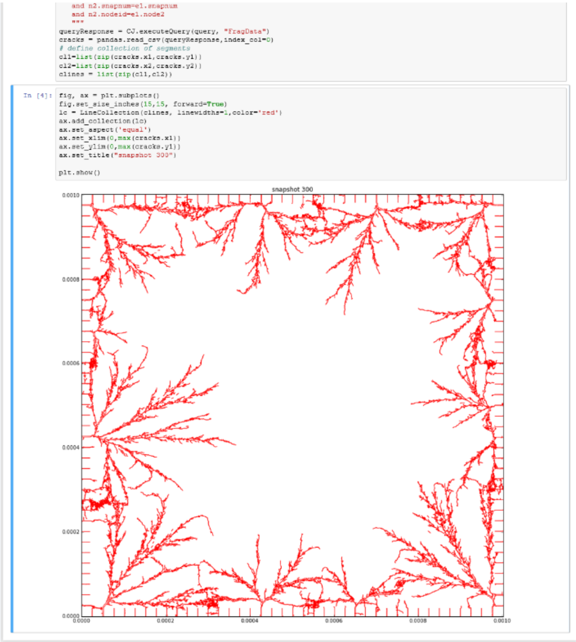
Now Avaliable: JupyterLab and Classical Jupyter images are now combined. Containers default to the classical interface and will remember the last interface used.

Containers

Created At	Name	Domain	Image	Status
2019-10-06 07:35:21.0	ADASS 2019 interactive GPU domain	GPU Interactive	TensorFlow	running
2019-10-02 08:36:11.0	heasoft	Interactive	Heasarc	stopped
2019-10-01 16:24:04.0	montage	Interactive	Montage	stopped
2019-09-30 14:44:39.0	ADASS2019-pytorch	Interactive	PyTorch	running
2019-07-09 17:45:08.0	indra (filedb)	Interactive	Python + R	stopped
2019-05-02 10:33:40.0	polybase and aris	Interactive	Python + R	stopped
2019-04-25 16:28:15.0	old english	Interactive	BeakerX	stopped
2019-03-19 11:46:22.0	WFIRST	Interactive	WFIRST-Archive SIT	stopped
2018-10-28 15:51:49.0	Sociology + III + ukbiobank	Interactive	Python + R	stopped

Create container





Collaborative data-driven science

SciServer

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

File Edit View Insert Cell Kernel Help

Code CellToolbar

```
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]:

```
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]:

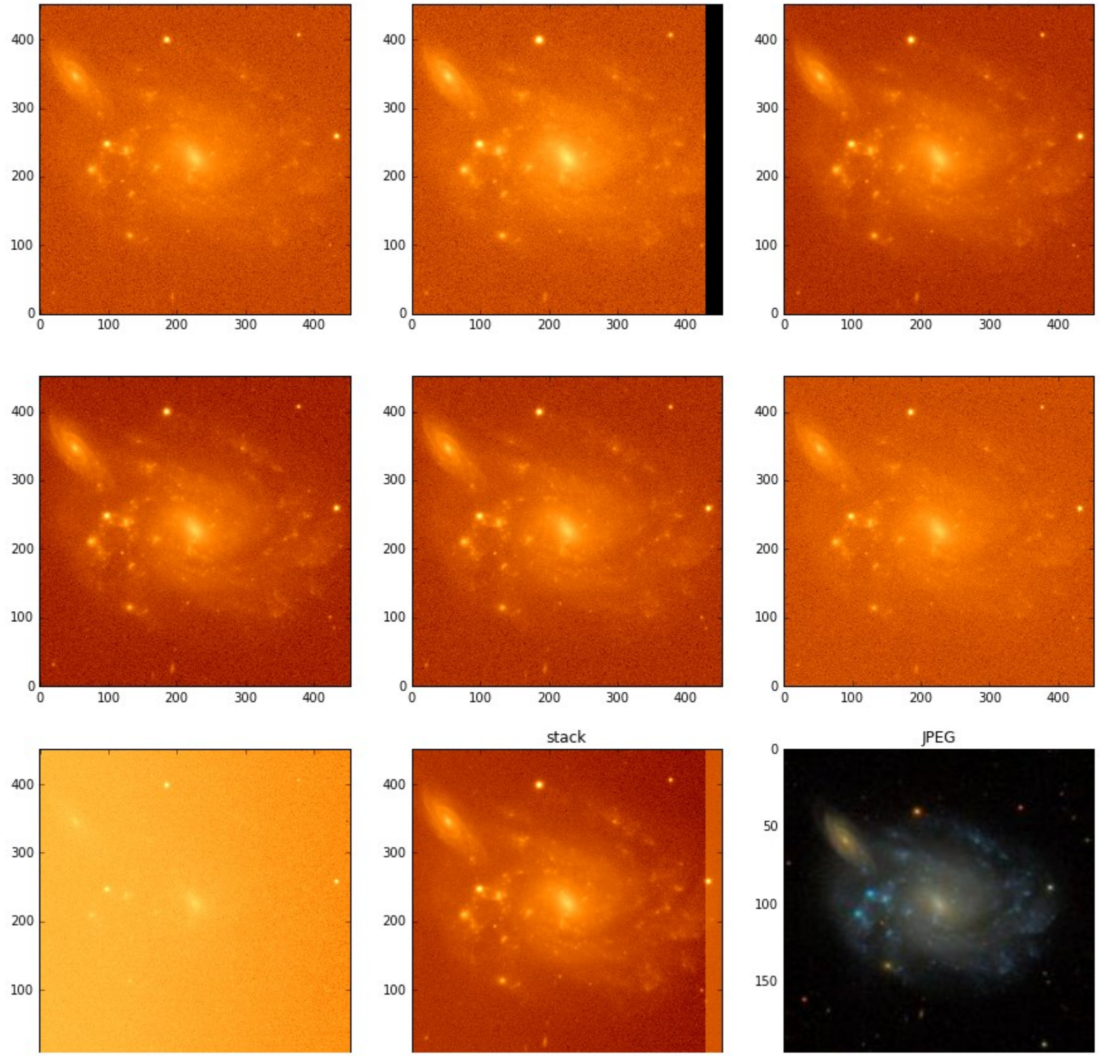
	head	match	matchcount	head_field	match_field	fits_g
0	8658194378960928809	8658194430499553320	57	8658194378960928768	8658194430499553280	http://das.sc
1	8658194378960928809	8658194477742948377	57	8658194378960928768	8658194477742948352	http://das.sc
2	8658194378960928809	8658194516375371821	57	8658194378960928768	8658194516375371776	http://das.sc
3	8658194378960928809	8658194585083510793	57	8658194378960928768	8658194585083510784	http://das.sc
4	8658194378960928809	8658194804163018771	57	8658194378960928768	8658194804163018752	http://das.sc
5	8658194378960928809	8658194954470752297	57	8658194378960928768	8658194954470752256	http://das.sc
6	8658194378960928809	8658195018907910161	57	8658194378960928768	8658195018907910144	http://das.sc
7	8658194378960928809	8658195044651040803	57	8658194378960928768	8658195044651040768	http://das.sc
8	8658194378960928809	8658195066151239724	57	8658194378960928768	8658195066151239680	http://das.sc
9	8658194378960928809	8658195113395748890	57	8658194378960928768	8658195113395748864	http://das.sc

jupyter 2. Stripe82-coadd-Copy1 Last Checkpoint: a minute ago (autosaved)

File Edit View Insert Cell Kernel Help Python 3

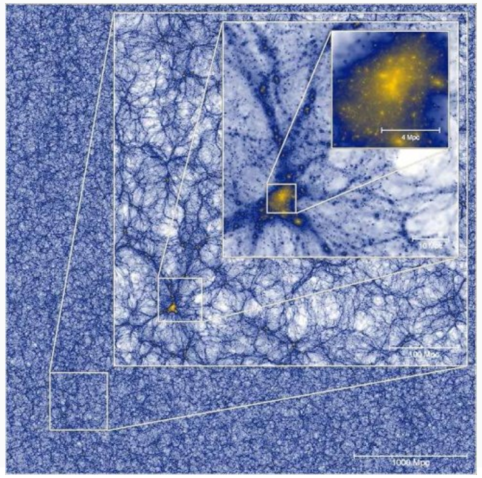
Code CellToolbar

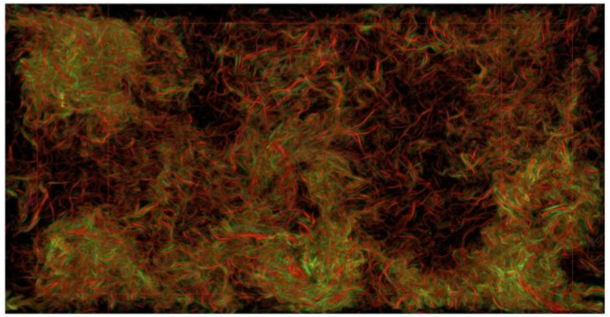
Out[17]: <matplotlib.text.Text at 0x7f2aa174a6d8>

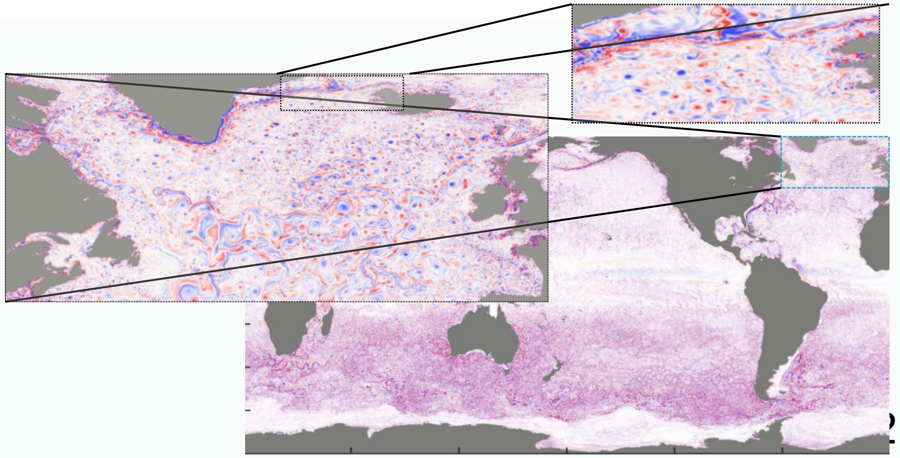


Simulations at SciServer

- Cosmology
  - Virgo Millennium suite ~40TB DB, ~60TB raw
  - Indra ~0.8PB
  - ApogeeFire DB ~29TB
  - Jason Hunt 2021 ~100TB
- Turbulence database, soon >2PB
- Ocean circulation, soon ~2PB







SciServer | Home | Files | Groups | Science Domains | Admin | Resources | gerard

Science Domains

- Cosmological Simulations
- Genomics

Cosmological Simulations

Join

Cosmological Simulations

If you join this science domain (by clicking the "join" button), you will have access to close to a peta-byte of data from a number of cosmological simulations.

The theoretical input for galaxy-scale physics and large-scale galaxy redshift surveys is provided by cosmological numerical simulations, which model the evolution of the Milky Way or the entire Universe. SciServer hosts several sets of these simulations, offering unified access and collaborative visualization tools for hundreds of TB of data. The unique advantage of accessing these simulations through SciServer is the ease with which they can be compared with real astronomical observations.

Some of the data sets made accessible in this domain are stored in relational databases and will be accessible through the CasJObs web page and the CasJobs modules in python and R. Others are accessible as files on data volumes. To use these data in SciServer Compute, create a container with the **Cosmological Simulations** compute image and mount the appropriate volumes.

Indra

Indra is a suite of large-volume cosmological N-body simulations, all with the same cosmology and different initial phases. Each simulation has 64 snapshots of particle data and halo catalogs and 505 time steps of Fourier-space density fields. The Indra data volume contains the full 750 TB dataset, and the Indra database contains the halo catalogs plus the Spatial3d library of search tools.

Notebooks giving examples of how to access Indra data are found in the Indra data volume and the indra-tools git repository.

- For an overview see: https://www.sciserver.org/datasets
- For detailed information see the data release paper: https://arxiv.org/abs/2101.03631
- Contact: Bridget Falck (bridget.falck@jhu.edu)

Virgo Consortium, Millennium suite of simulations

This science domain also gives access to a variety of simulations and related products produced by the Virgo Consortium for Cosmological Supercomputer Simulations. In particular SciServer contains a copy of the Millennium Database, which contains halo catalogues and catalogues of galaxies simulated using semi-analytical galaxy formation algorithms.

SciServer extends the functionality of that site by providing online analysis capabilities in Jupyter notebooks. But in particular SciServer allows user to access the underlying raw simulation data, i.e. particles for the N-body simulations. Notebooks in the Getting Started data volume give examples how to access these.

See also:

- For more details of the Millennium simulations and their usage by the community see https://wwwmpa.mpa-garching.mpg.de/millennium/
- For more details of the Millennium Database please see the original web site.

References

- Millennium Simulation: Springel etal 2005
- Millennium Database: Lemson & the Virgo Consortium 2006

Collaborative dat

jupyter Query Merger Trees Last Checkpoint: 11/08/2019 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3.8 (py38)

Run Code

In [1]:

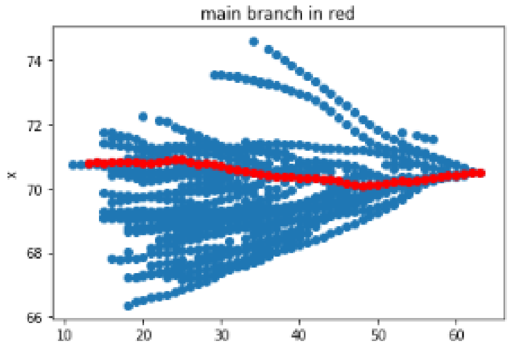
```
import SciServer.CasJobs as cj
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:

```
# select a random halo at z=0 (snapnum=63) and mass between 500 and 5000x10^10 Msun/h
# then find merger tree rooted in that halo
sql="""
with descendants as (
select top 1 haloid , lastprogenitorid, mainleafid, m_crit200, d.np
  from MR d
 where snapnum=63                        -- reshift=0
   and firsthaloinfofgroupid=d.haloid  -- centers of FOF groups
   and m_crit200 between 500 and 5000
 order by newid()                        -- random sub sample
)
select d.haloid as d_id, d.m_crit200 as d_mass, d.np as d_np
,      p.haloId,p.x,p.y,p.z,p.snapnum,p.mainLeafId,p.halfMassRadius
  from descendants d
      inner join MR p
        on p.haloid between d.haloid and d.lastprogenitorid
order by d.haloid, p.haloid
"""
tree=cj.executeQuery(sql,"MPAHaloTrees")  # send query to the MPAHaloTrees database context
```

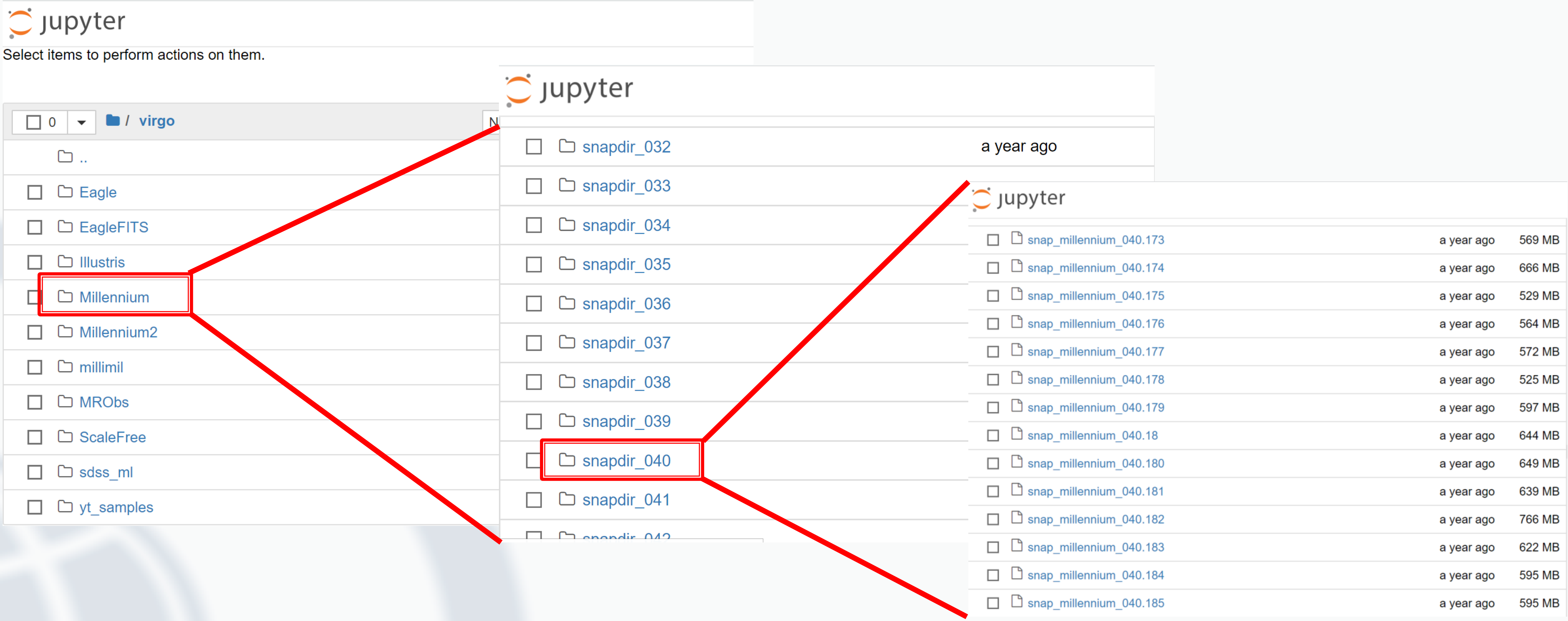
In [4]:

```
X='snapnum'
Y='x'
i=0
#for d_id,tree in gdf:
i1,i2=tree[['haloId','mainLeafId']].iloc[0]
tree.set_index('haloId',inplace=True)
main=tree.loc[i1:i2]
plt.scatter(tree[X],tree[Y])
plt.scatter(main[X],main[Y],color='red')
plt.xlabel(X)
plt.ylabel(Y)
plt.title('main branch in red');
```



rver

Raw data, accessible from inside and outside the database



SciServer

jupyter CosmoUC_5_PlotHalo Last Checkpoint: 11 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Notebook saved

Code CellToolbar

Dark-matter Halos in a Cosmological Simulation

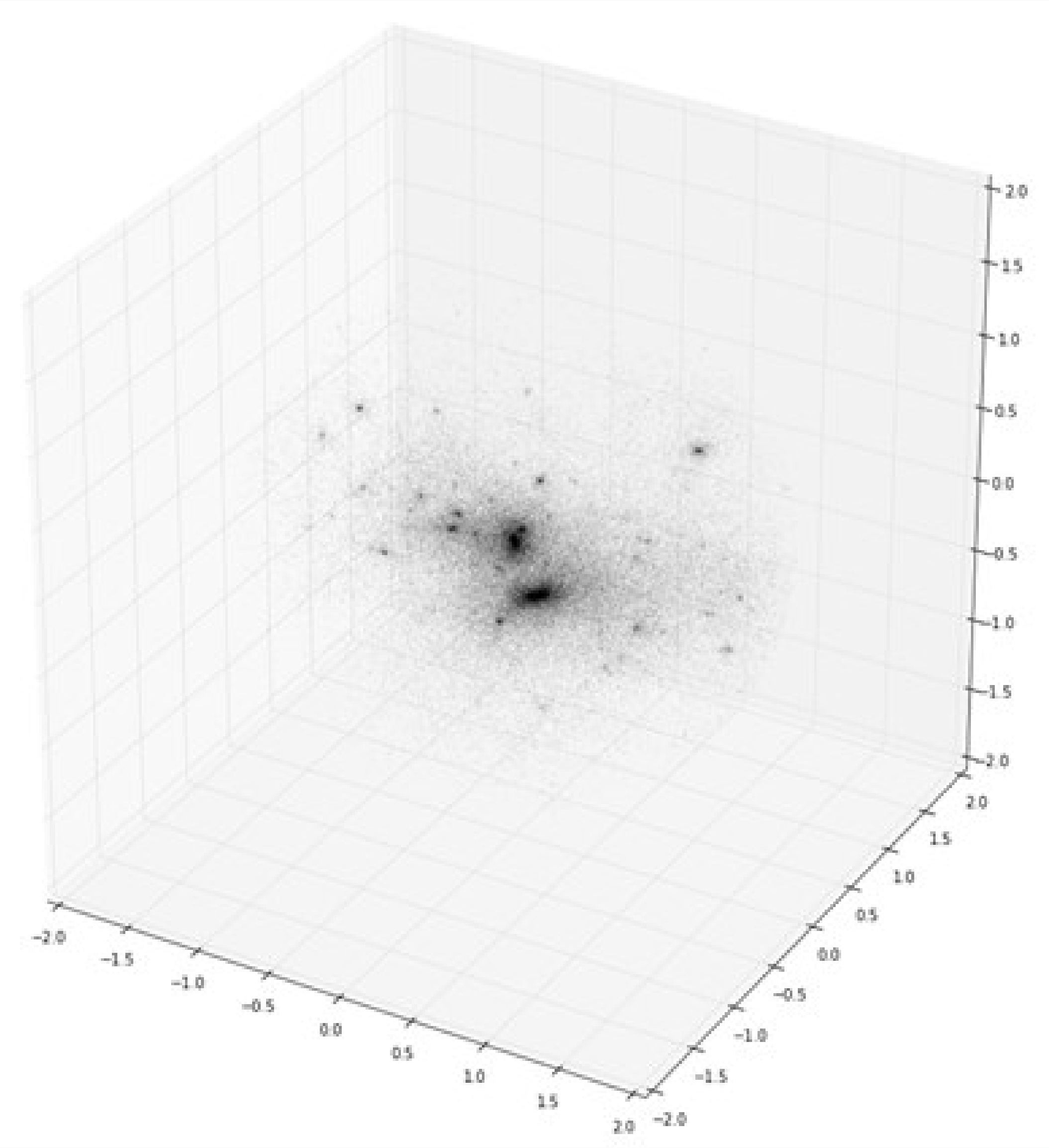
In [3]:
```
import SciServer.LoginPortal as Login
token = Login.getToken()
import SciServer.CasJobs
import pandas
import tables
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
```

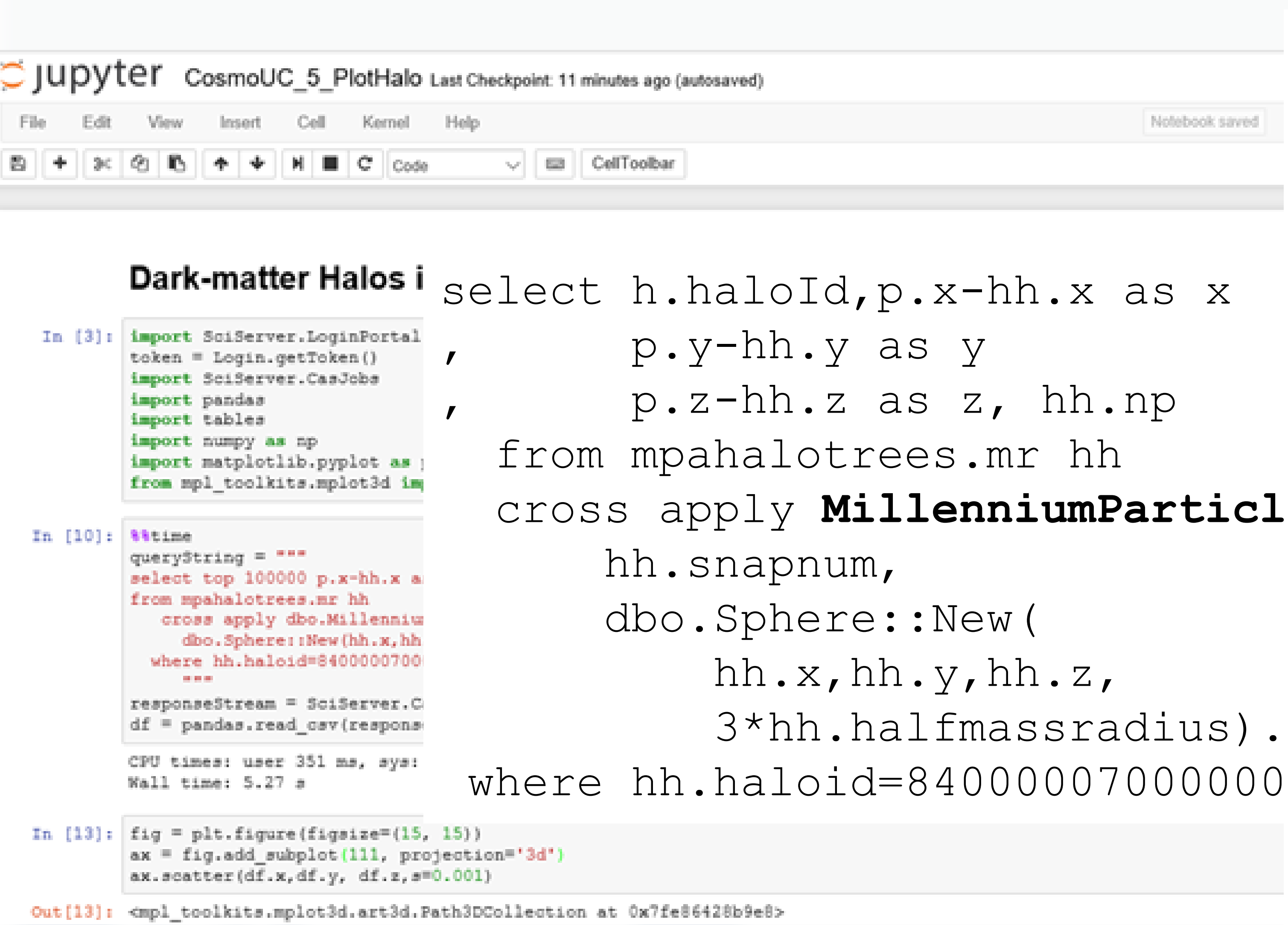
In [10]:
```
%%time
queryString = """
select top 100000 p.x-hh.x as x,p.y-hh.y as y,p.z-hh.z as z
from mpahalotrees.mr hh
   cross apply dbo.MillenniumParticles(hh.snapnum,
     dbo.Sphere::New(hh.x,hh.y,hh.z,3*hh.halfmassradius).ToString()) p
  where hh.haloid=84000007000000 order by newid()
     """
responseStream = SciServer.CasJobs.executeQuery(queryString, token=token,context="SimulationDB")
df = pandas.read_csv(responseStream, index_col=None)
```

```
CPU times: user 351 ms, sys: 184 ms, total: 535 ms
Wall time: 5.27 s
```

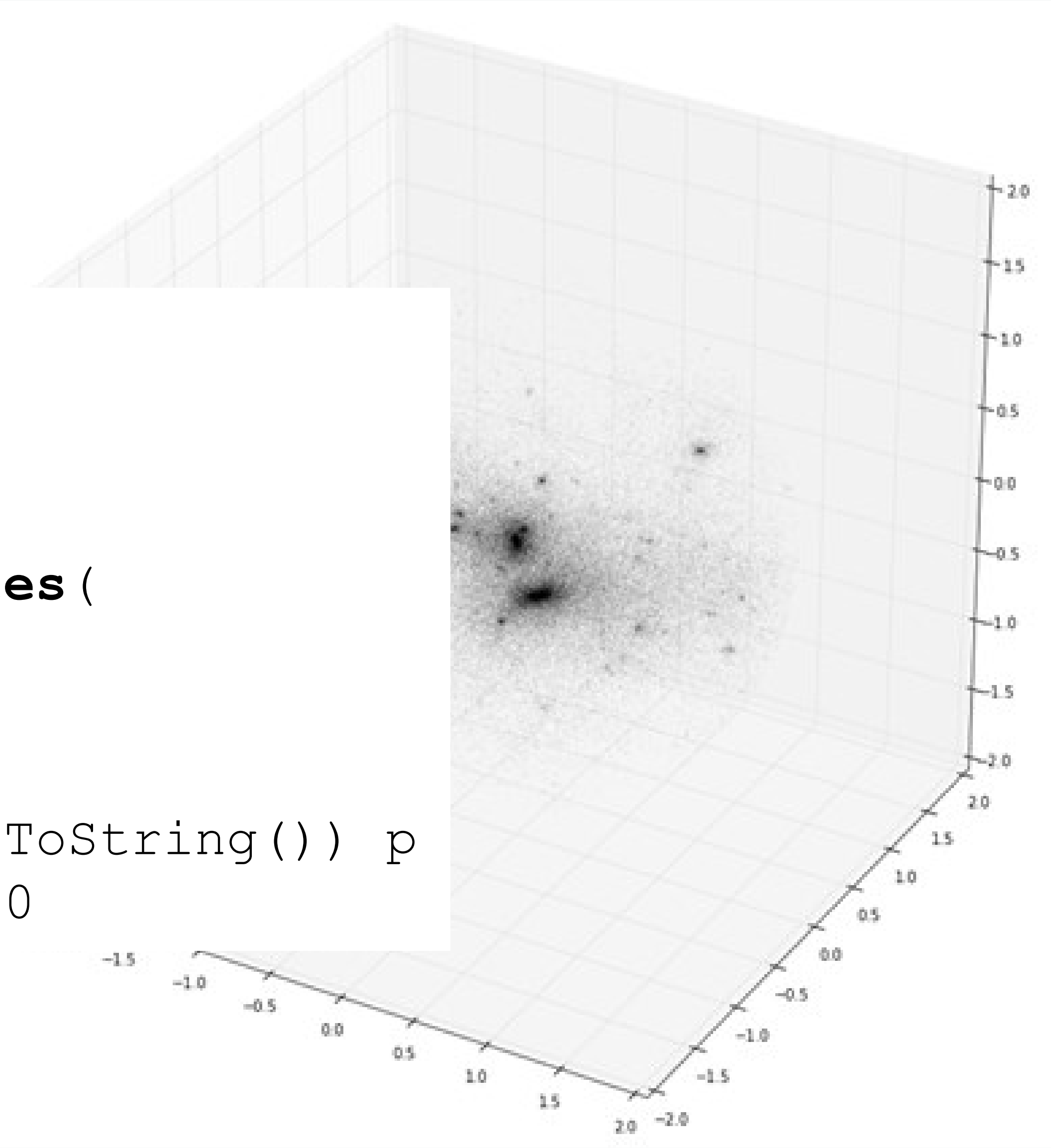
In [13]:
```
fig = plt.figure(figsize=(15, 15))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.x,df.y, df.z,s=0.001)
```

Out[13]: <mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7fe86428b9e8>






```
select h.haloId,p.x-hh.x as x
,       p.y-hh.y as y
,       p.z-hh.z as z, hh.np
  from mpahalotrees.mr hh
  cross apply MillenniumParticles(
      hh.snapnum,
      dbo.Sphere::New(
          hh.x,hh.y,hh.z,
          3*hh.halfmassradius).ToString()) p
 where hh.haloid=840000070000000
```

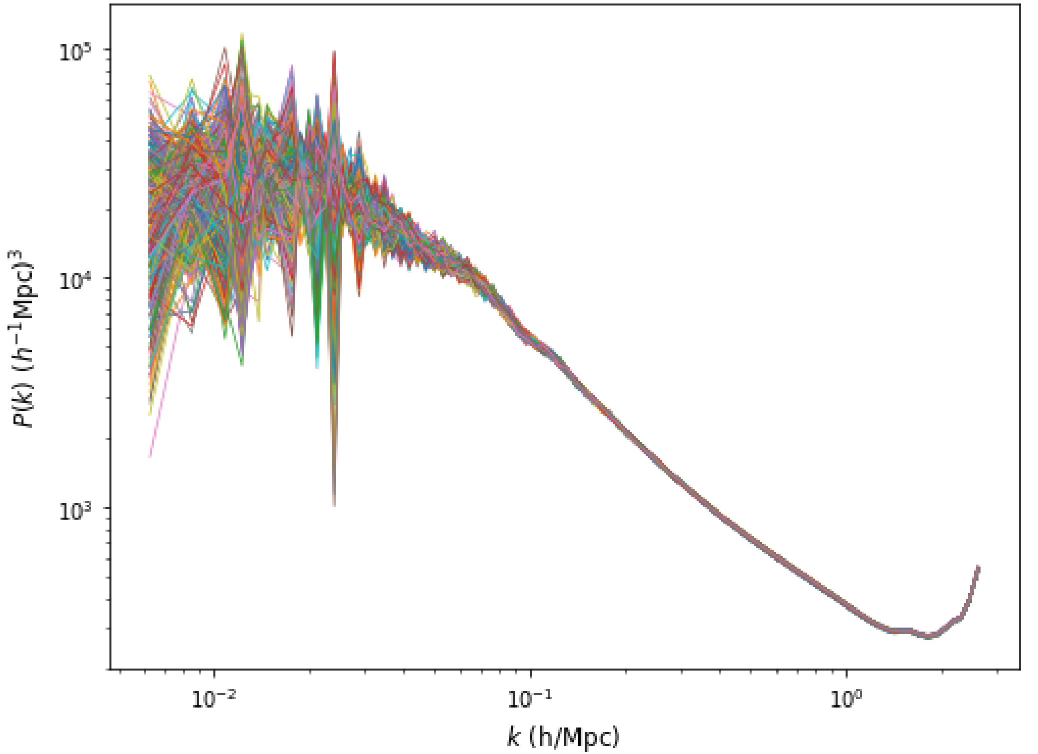


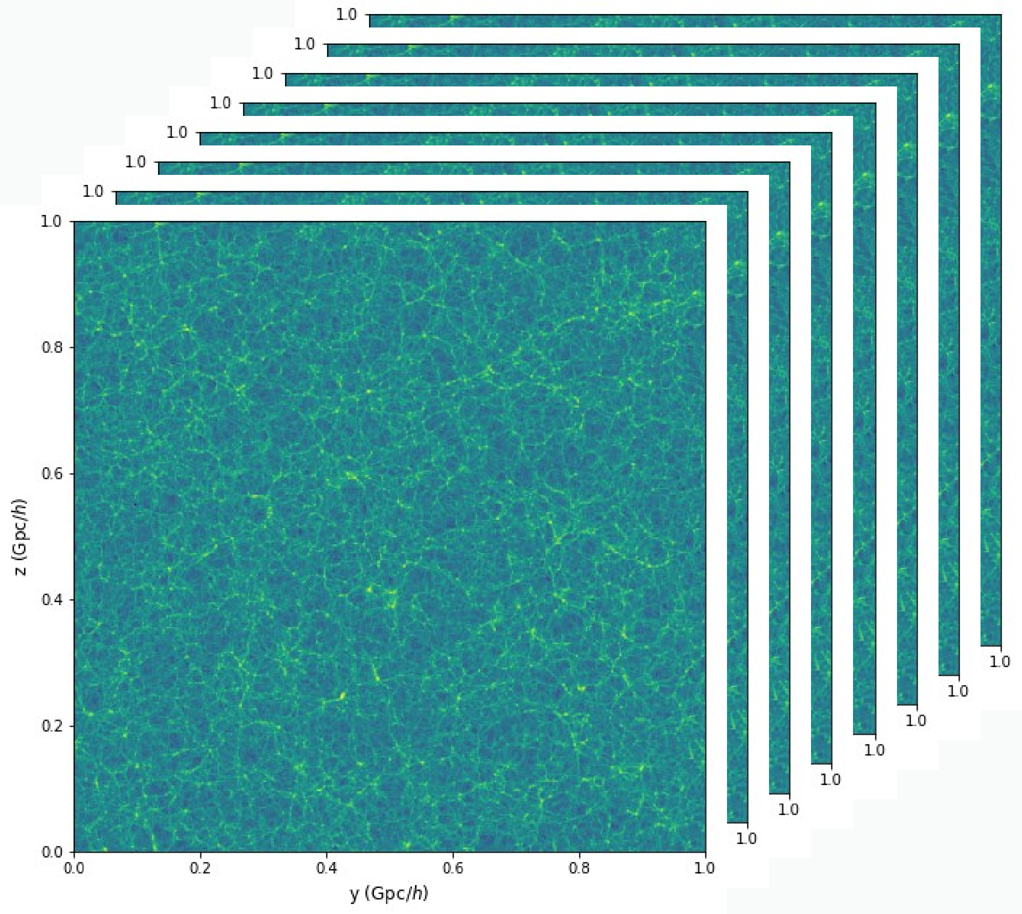
Indra: a public computationally accessible suite of cosmological N-body simulations

Falck etal 2021, https://ui.adsabs.harvard.edu/abs/2021MNRAS.506.2659F/abstract

Power spectra with DASK

448 simulations, 1 billion particles each.
Same initial conditions, different seed.
448 Cloud-In-Cell density grids created and FFT-ed in 2 hours, using 8 DASK workers on distributed file system

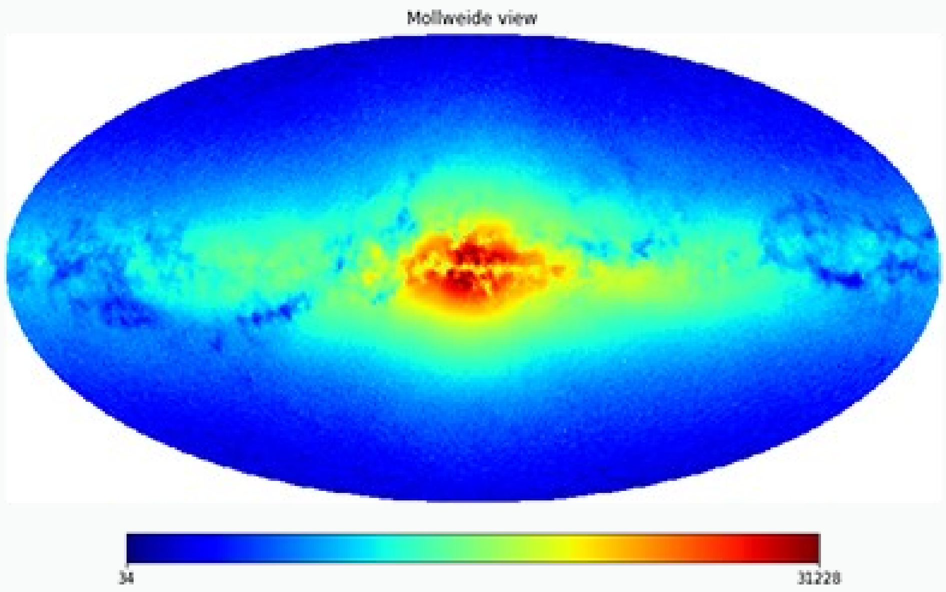




APOGEE-centric Ananke Simulations in a SciServer SQL Database

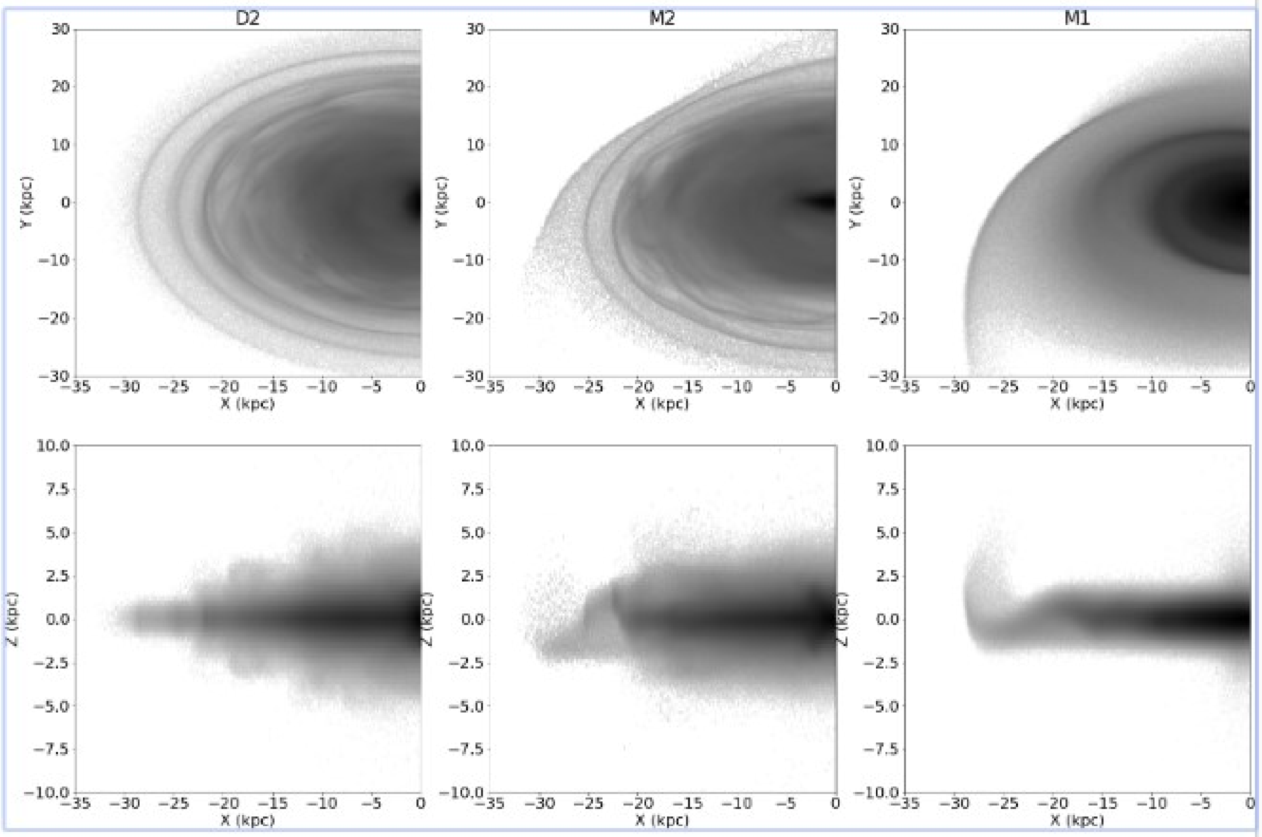
Beaton etal 2022, https://ui.adsabs.harvard.edu/abs/2022RNAAS...6..125B

- ELT process on Spark:
  - Globus transfer of data
  - Spark/HDFS used for transforming FITS→parquet
  - Addition of special columns:
    - htm20Id, heal20Id, positions in different coordinate systems, …
  - MS libraries used for loading into RDB column store
  - Registration in SciServer for access through CasJobs and from within Jupyter notebooks

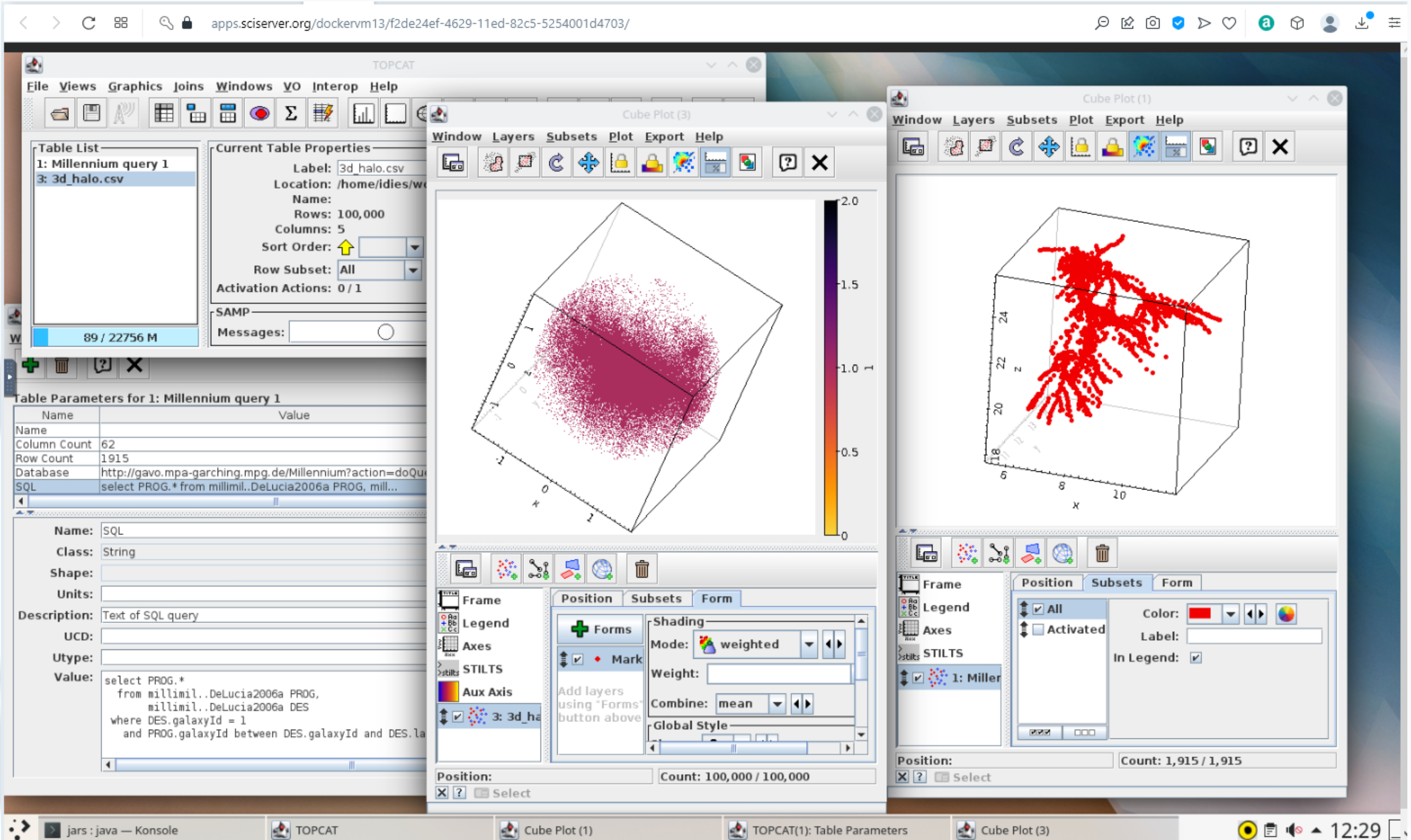


Resolving local and global kinematic signatures of satellite mergers with billion particle simulations Hunt etal 2021, https://ui.adsabs.harvard.edu/abs/2021MNRAS.508.1459H/abstract

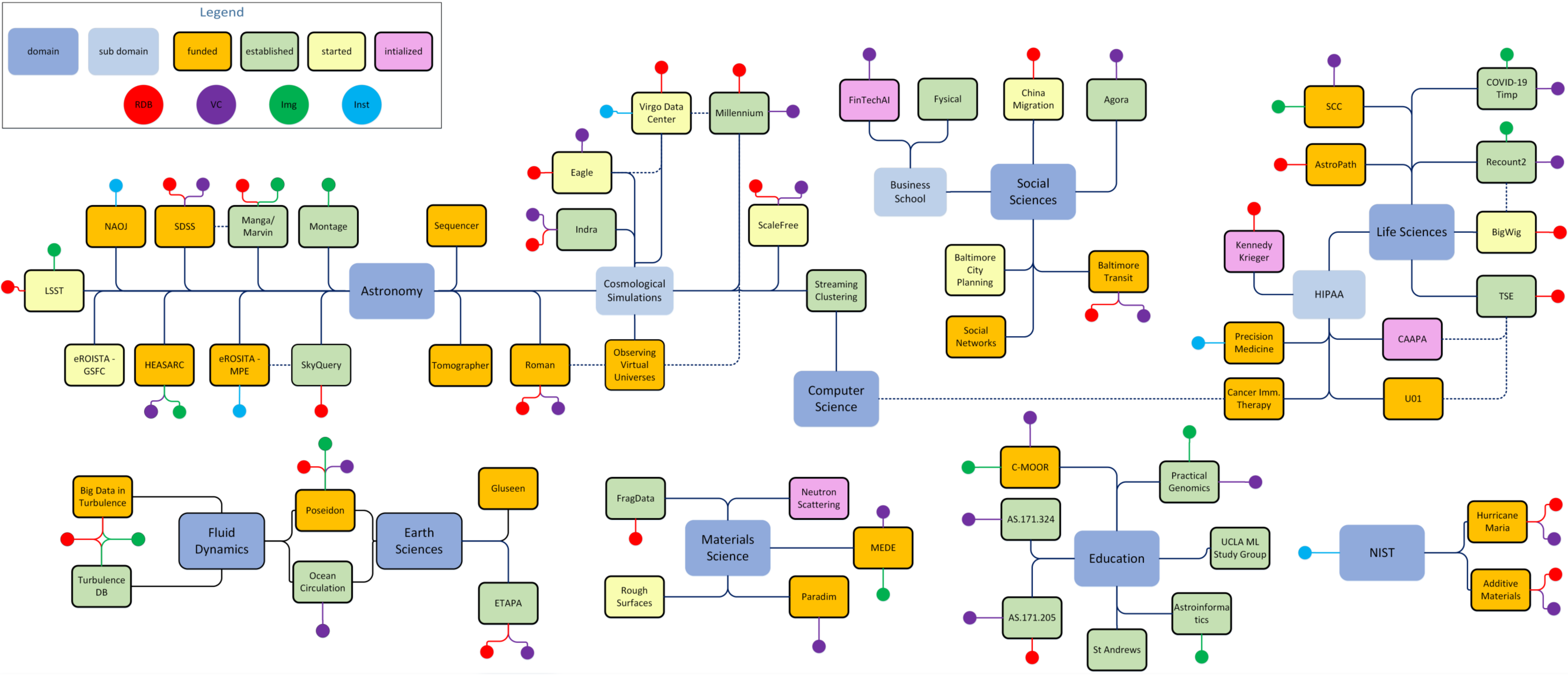
- Data transfer through Globus tasks scripted in Jupyter notebooks on SciServer to distribute files over 12x3 storage volumes of *FileDB cluster*
- Analysis on sciserver in progress (C. Fillion)



VNC KDE compute image: linux desktop

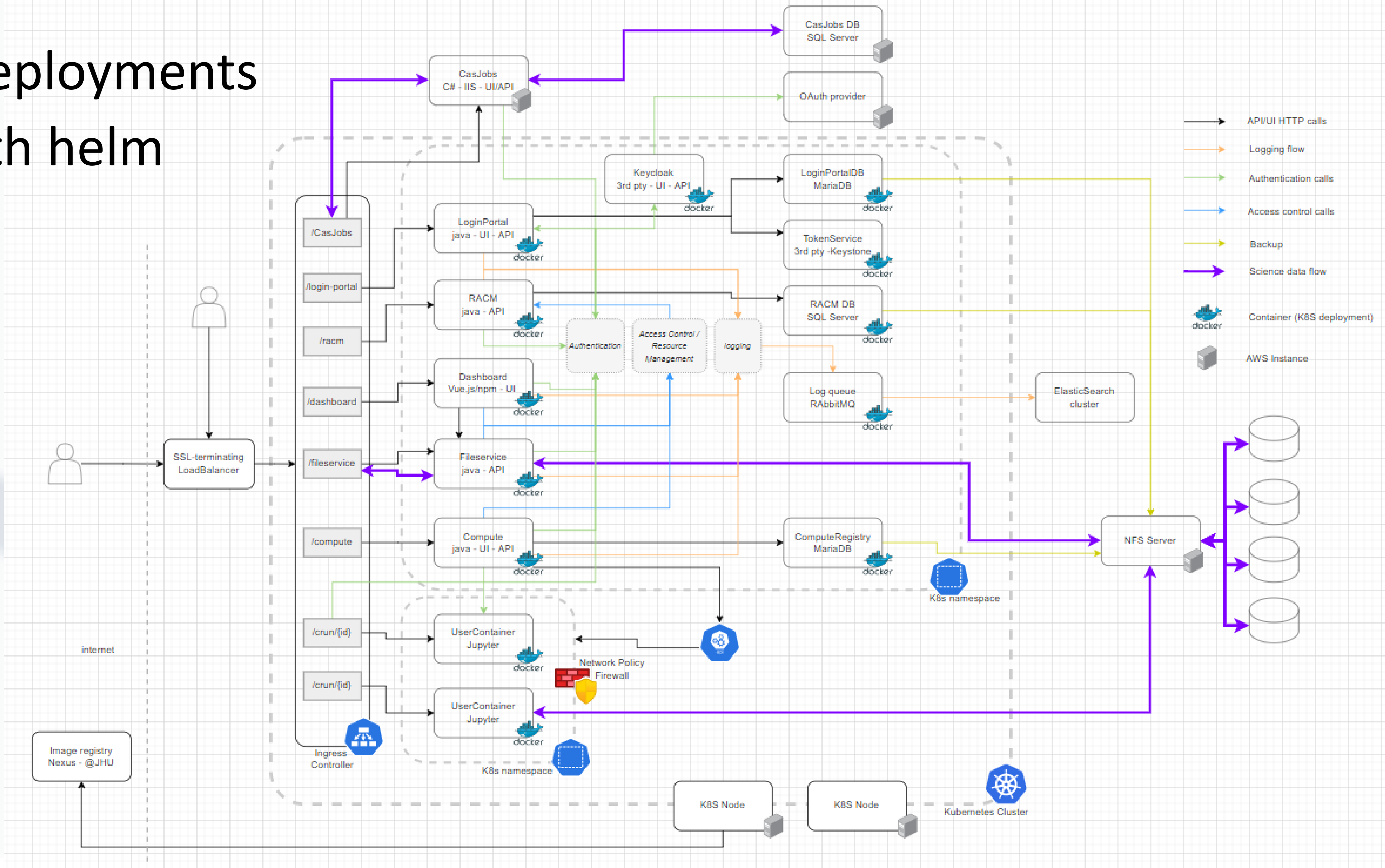


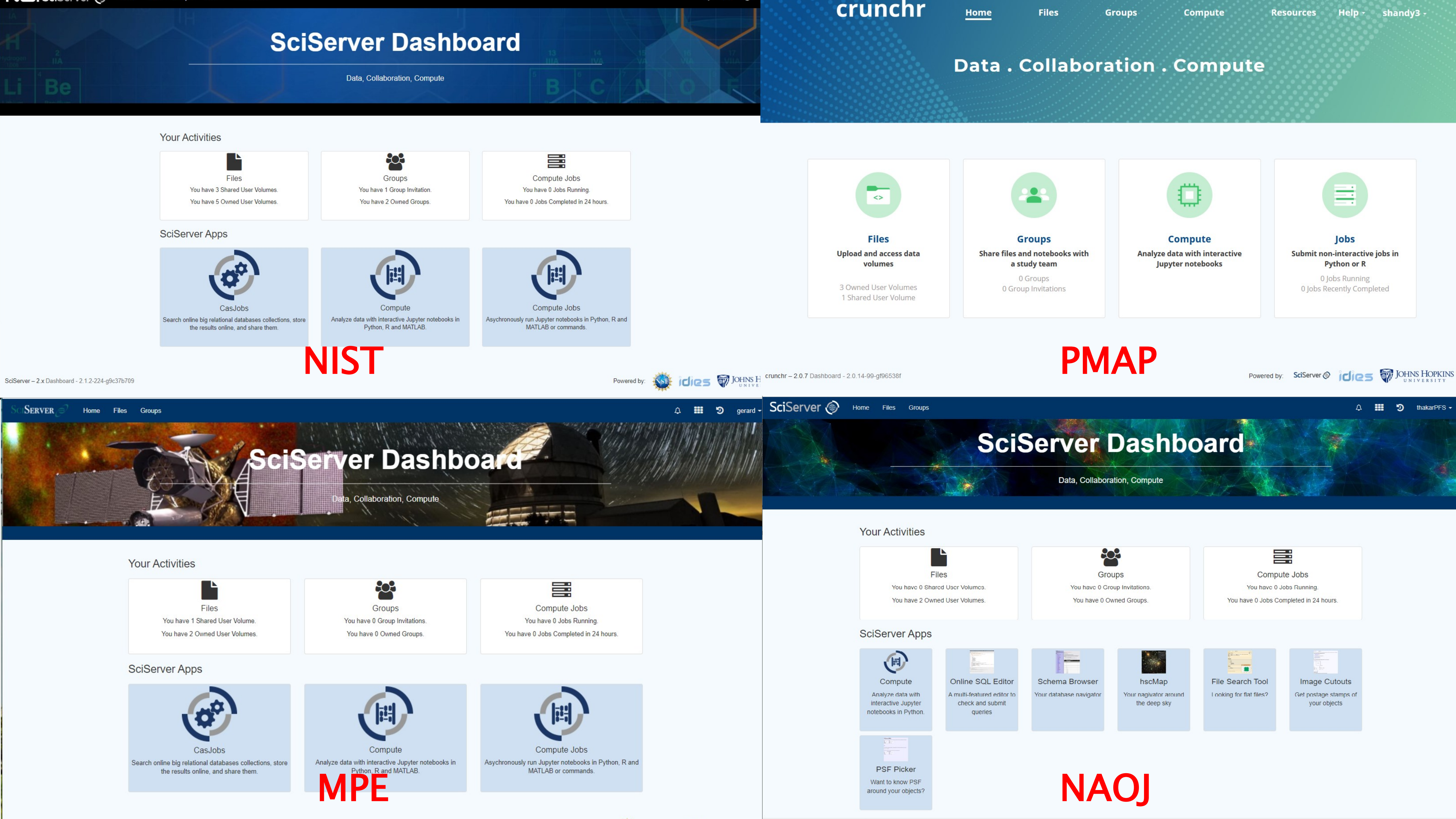
Beyond astrophysics only



SciServer

- External deployments
- On K8S with helm




SciServer Dashboard
Data, Collaboration, Compute
Your Activities
Files
You have 3 Shared User Volumes.
You have 5 Owned User Volumes.
Groups
You have 1 Group Invitation.
You have 2 Owned Groups.
Compute Jobs
You have 0 Jobs Running.
You have 0 Jobs Completed in 24 hours.
SciServer Apps
CasJobs
Search online big relational databases collections, store the results online, and share them.
Compute
Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.
Compute Jobs
Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.
NIST
SciServer – 2.x Dashboard - 2.1.2-224-g9c37b709
Powered by:
crunchr
Home
Files
Groups
Compute
Resources
Help
shandy3
Data . Collaboration . Compute
Files
Upload and access data volumes
3 Owned User Volumes
1 Shared User Volume
Groups
Share files and notebooks with a study team
0 Groups
0 Group Invitations
Compute
Analyze data with interactive Jupyter notebooks
Jobs
Submit non-interactive jobs in Python or R
0 Jobs Running
0 Jobs Recently Completed
PMAP
crunchr – 2.0.7 Dashboard - 2.0.14-99-gf96538f
Powered by:
Home
Files
Groups
gerard
SciServer Dashboard
Data, Collaboration, Compute
Your Activities
Files
You have 1 Shared User Volume.
You have 2 Owned User Volumes.
Groups
You have 0 Group Invitations.
You have 0 Owned Groups.
Compute Jobs
You have 0 Jobs Running.
You have 0 Jobs Completed in 24 hours.
SciServer Apps
CasJobs
Search online big relational databases collections, store the results online, and share them.
Compute
Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.
Compute Jobs
Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.
MPE
SciServer
Home
Files
Groups
thakarPFS
SciServer Dashboard
Data, Collaboration, Compute
Your Activities
Files
You have 0 Shared User Volumes.
You have 2 Owned User Volumes.
Groups
You have 0 Group Invitations.
You have 0 Owned Groups.
Compute Jobs
You have 0 Jobs Running.
You have 0 Jobs Completed in 24 hours.
SciServer Apps
Compute
Analyze data with interactive Jupyter notebooks in Python.
Online SQL Editor
A multi-featured editor to check and submit queries
Schema Browser
Your database navigator
hscMap
Your nagivator around the deep sky
File Search Tool
Looking for flat files?
Image Cutouts
Get postage stamps of your objects
PSF Picker
Want to know PSF around your objects?
NAOJ

SciServer, a Scalable Data Integrator

- Difficult to aggregate large, geographically distributed data sets:
  - joint analysis requires co-location.
  - SciServer facilitates creating ETL pipelines
- Joint analysis requires integration: aggregate data from various sources in a common context
  - A Science Platform like SciServer provides such a context
  - It allows users with diverse skills to collaborate on a single data set
- Most frequent mistake: trying to create the “mother of all databases”
  - Building ontologies and data models is hard
  - We learned an enormous amount during the Virtual Observatory project
- Real life uses require interactive exploration before big analysis

The SciServer philosophy:

- Create Data Contexts, each possibly with their own data model and ontology, self documenting
- These are secure and read-only, under access control
- User get their own databases/user volumes and resources to create value added results
- These can be shared at will with authenticated users at owner's discretion
- We can bring in new datasets in isolation very quickly
- **Reproducible science on integrated system**

Thank you

Registration is free at

https://www.sciserver.org

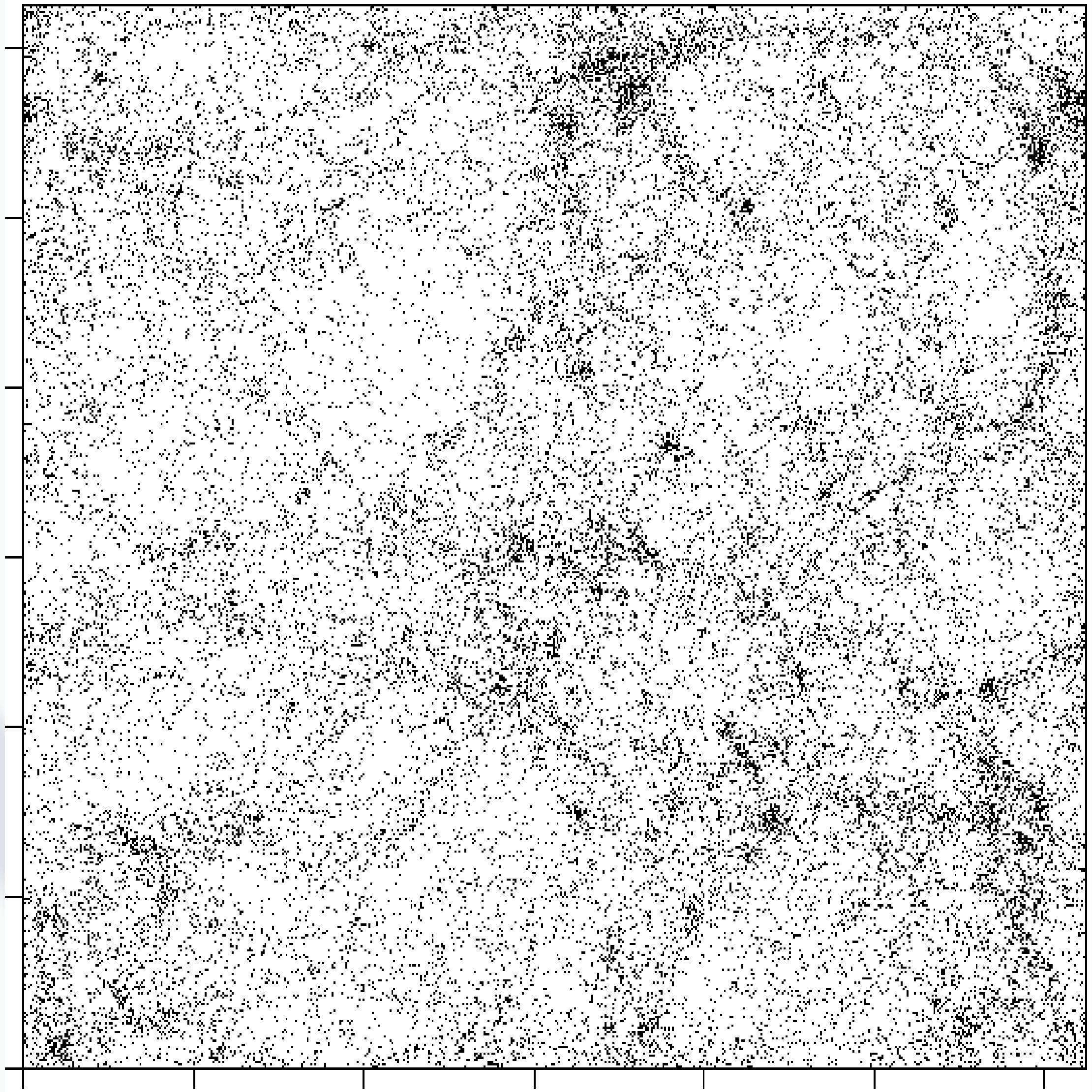
https://apps.sciserver.org

Appendix:
Efficiently querying 3D sub-volumes

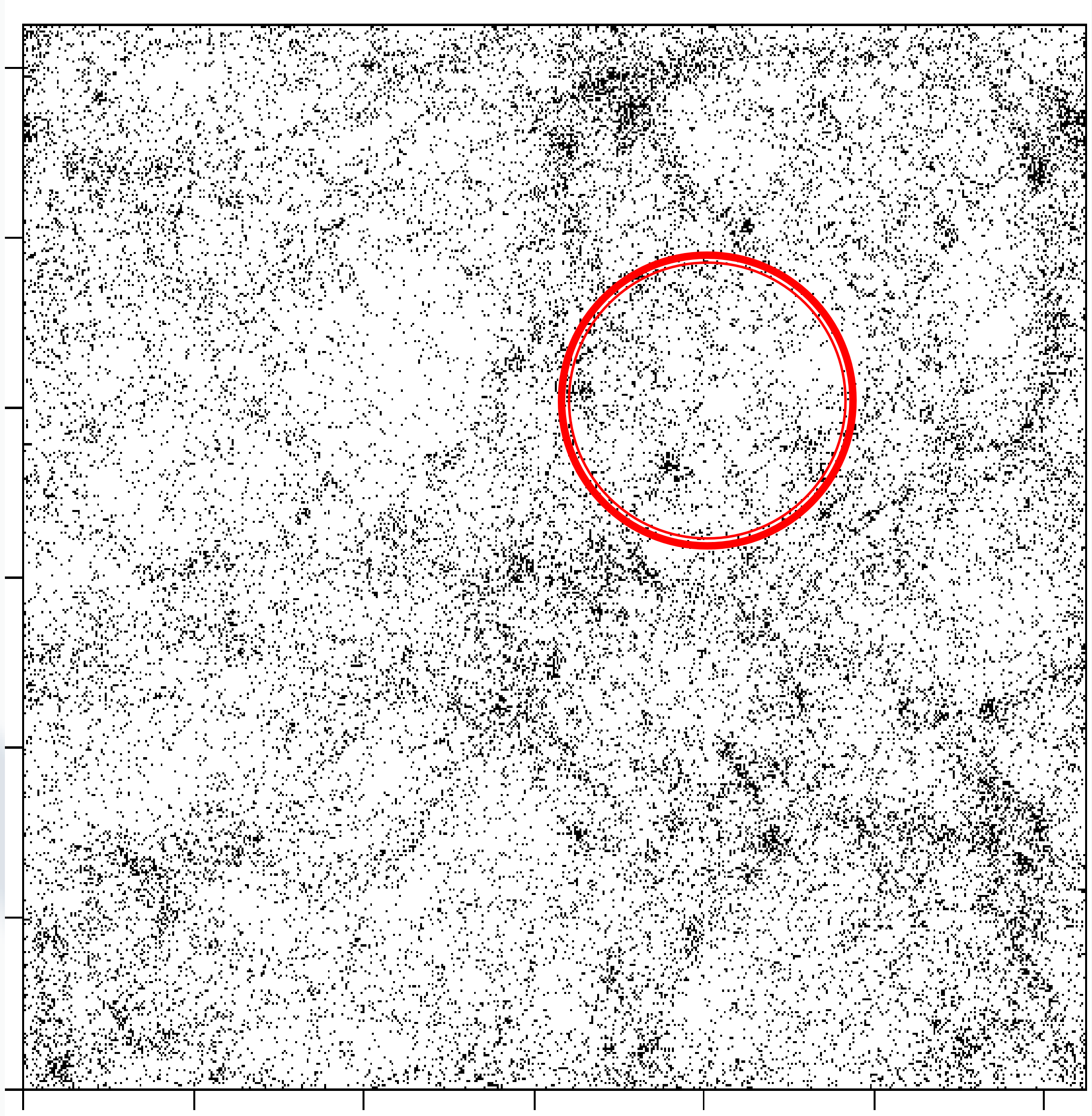
Implementing a General Spatial Indexing Library for Relational Databases of Large Numerical Simulations

Lemson, Budavari & Szalay, 2012

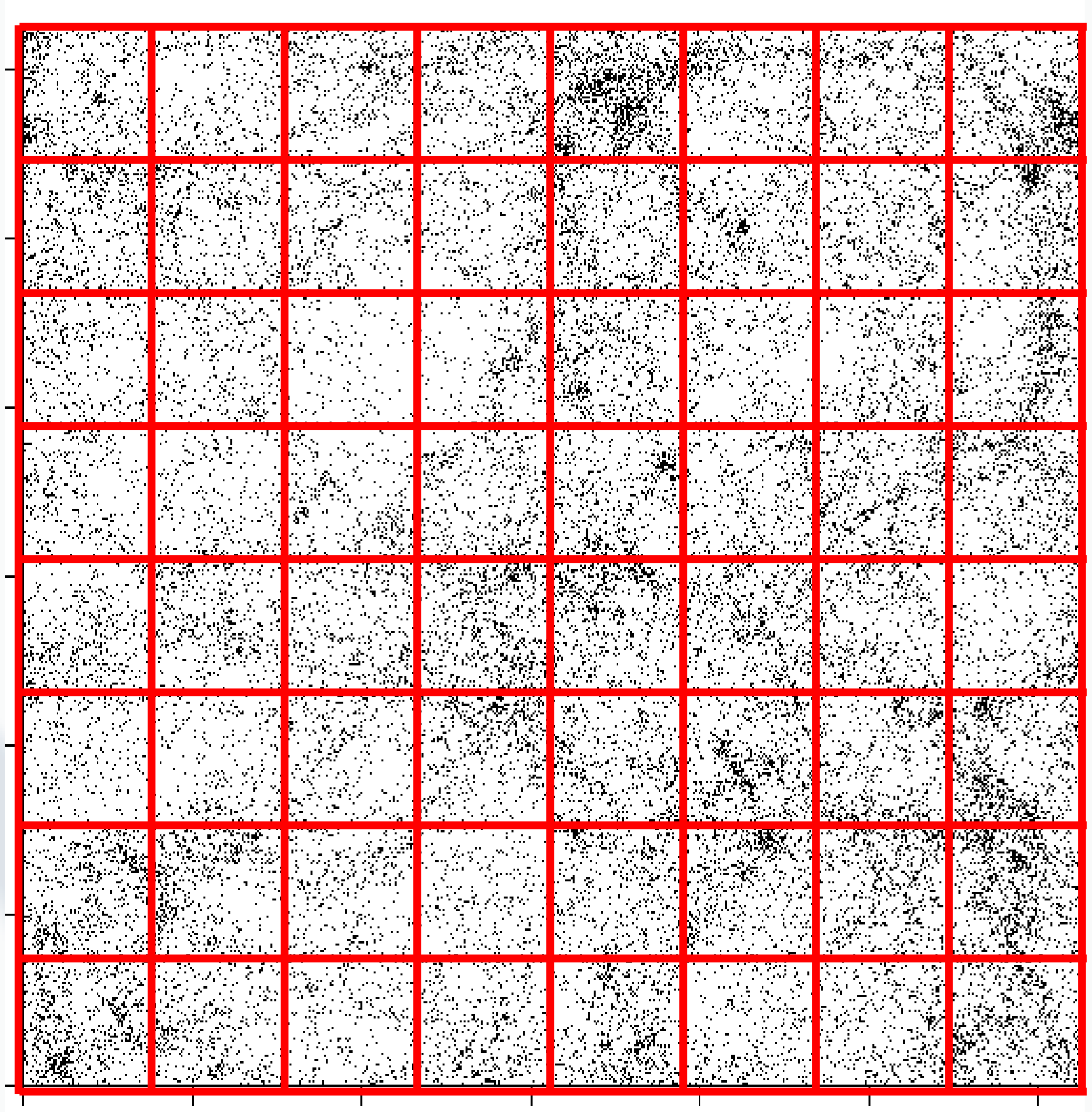
https://link.springer.com/chapter/10.1007/978-3-642-22351-8_34



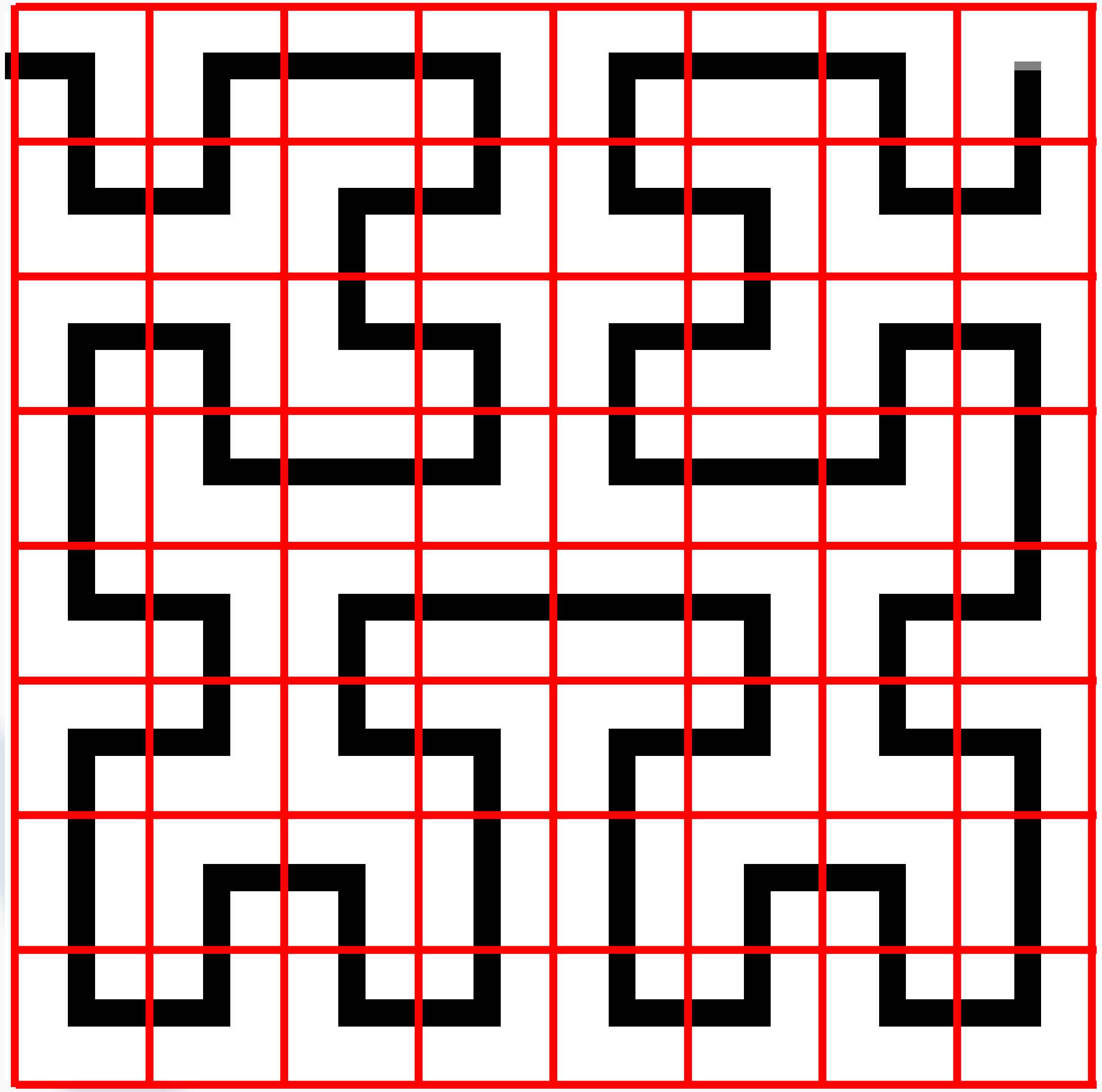
Find particles in some sub-volume



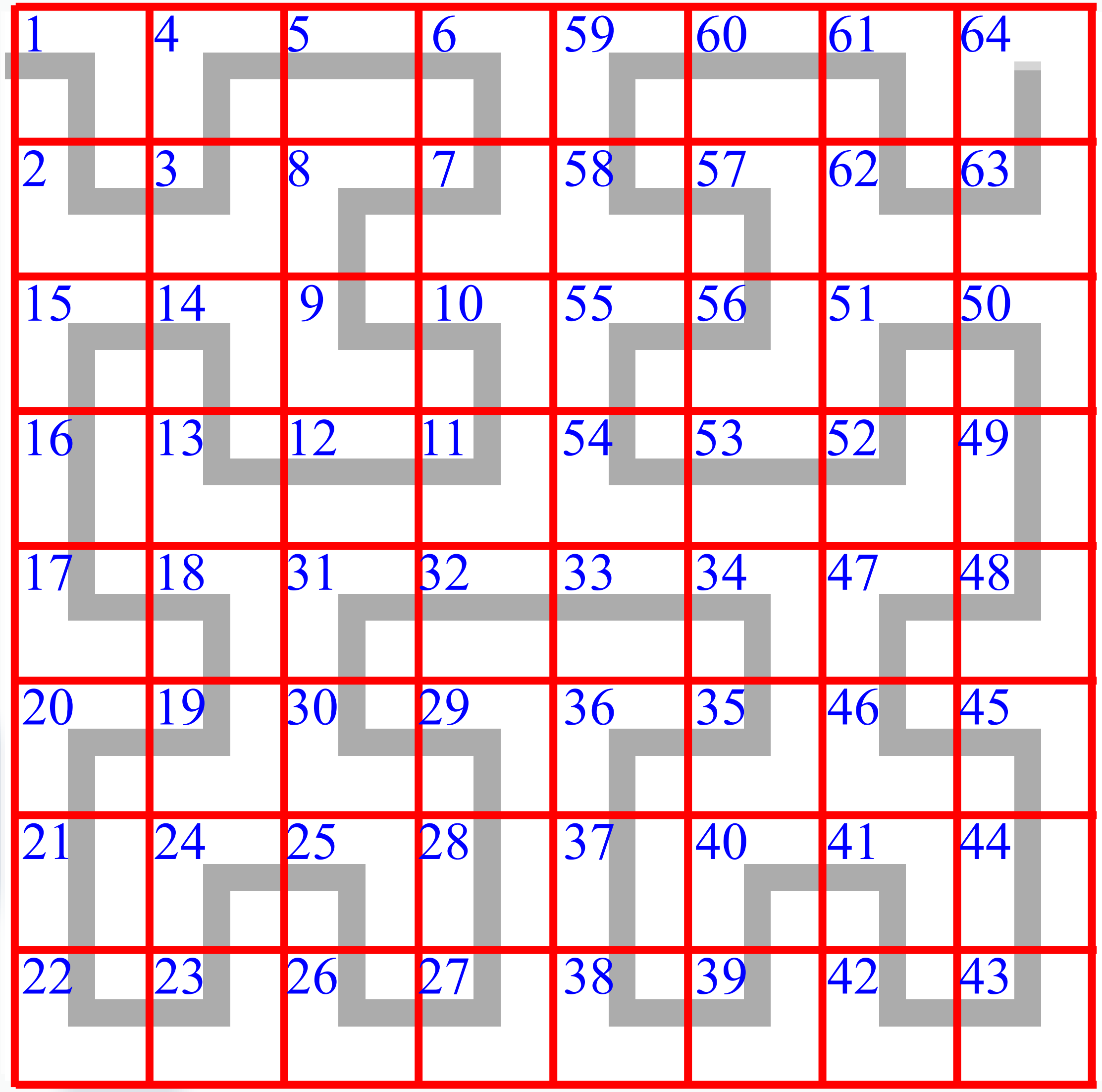
Divide simulation volume in regular grid



Index cells using space filling curve



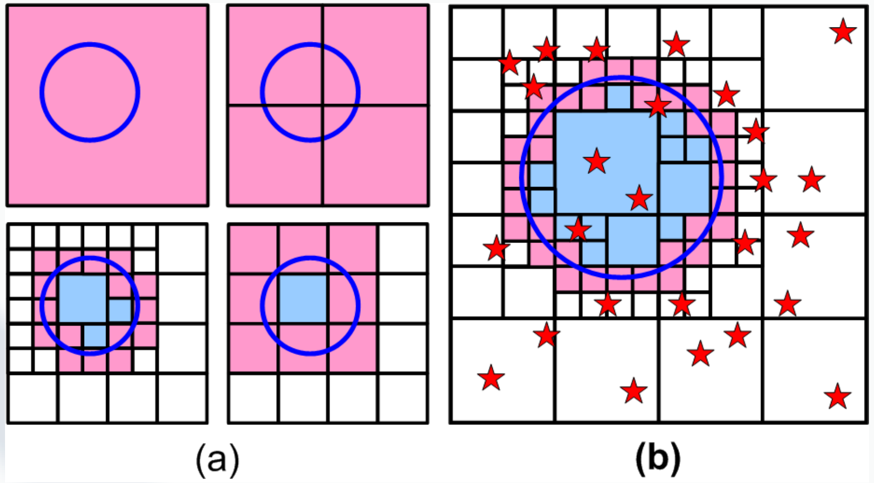
Sort data by index, in DB or files

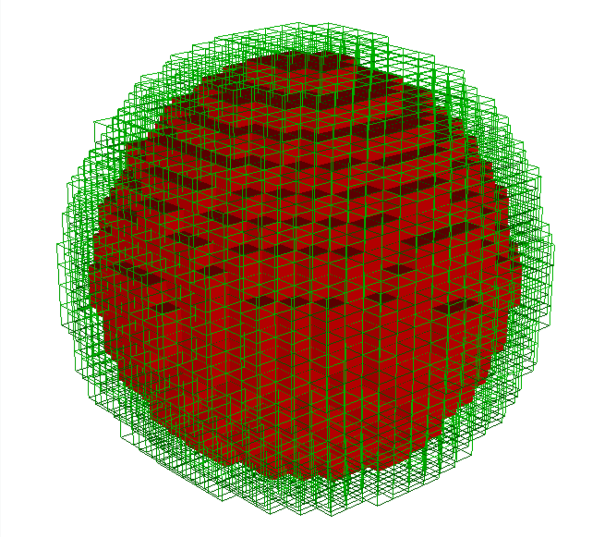


Lemson, Budavari & Szalay (2011)

https://link.springer.com/chapter/10.1007/978-3-642-22351-8_34

- Divide simulation volume in regular grid
- Index using space filling curve (Peano-Hilbert, Morton)
- Calculate overlap space filling curve with query volume
  - Iterate from root volume down, stopping at fully contained boxes





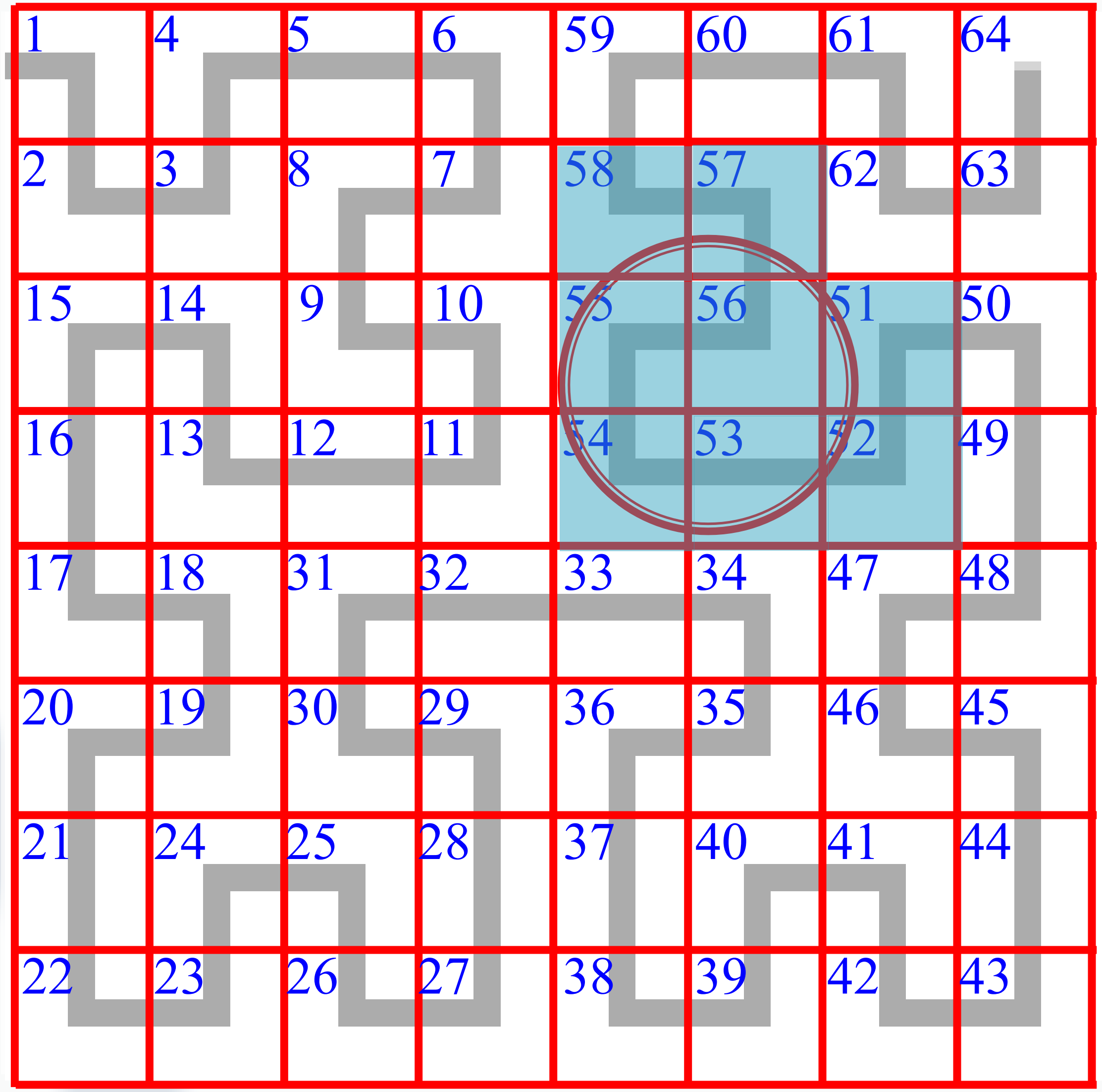
- Find particles in overlap ranges,
  - Only for those filter further on exact 3D volume
- Execute as table-valued function from database

Space-filling ordering co-locates near-by data better

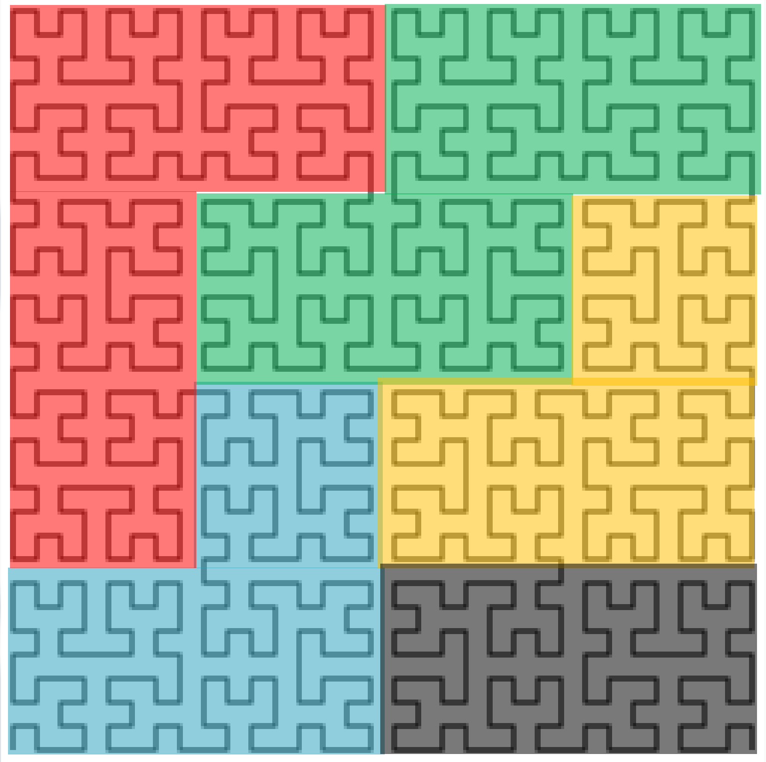
51 52 53 54 55 56 57

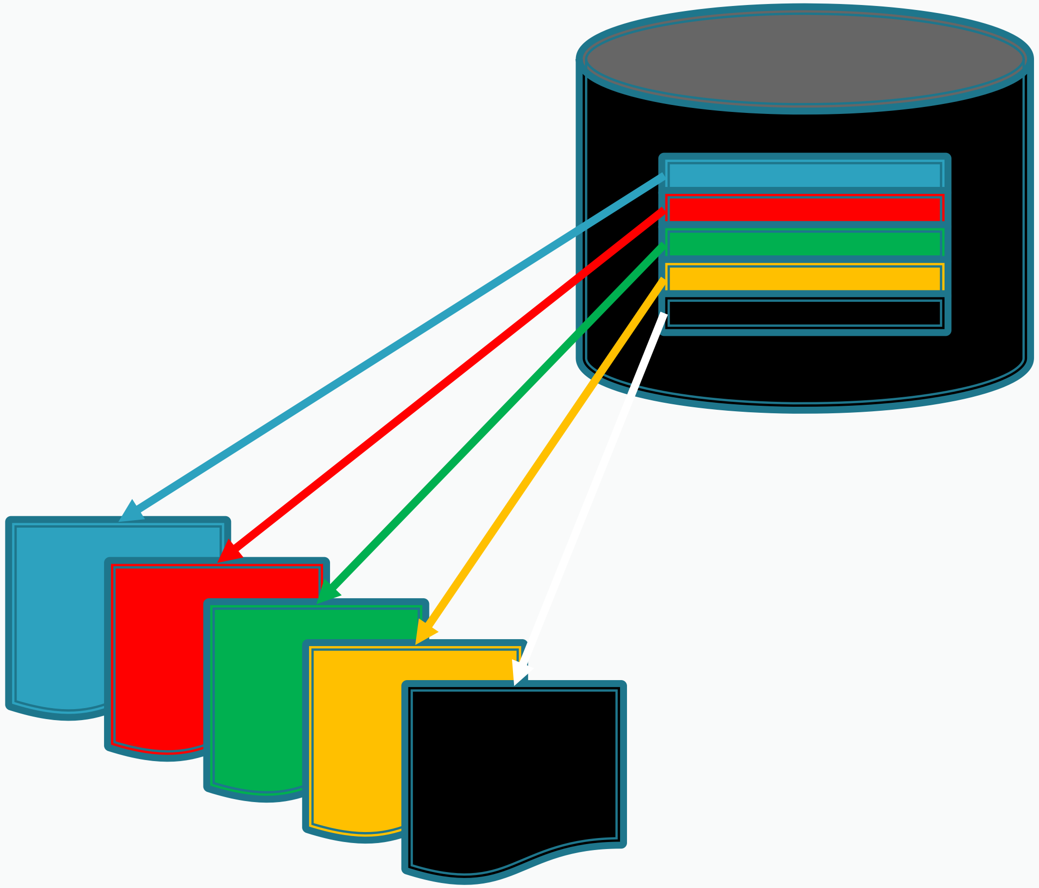
vs

15 16
23 24 25
31 32 33

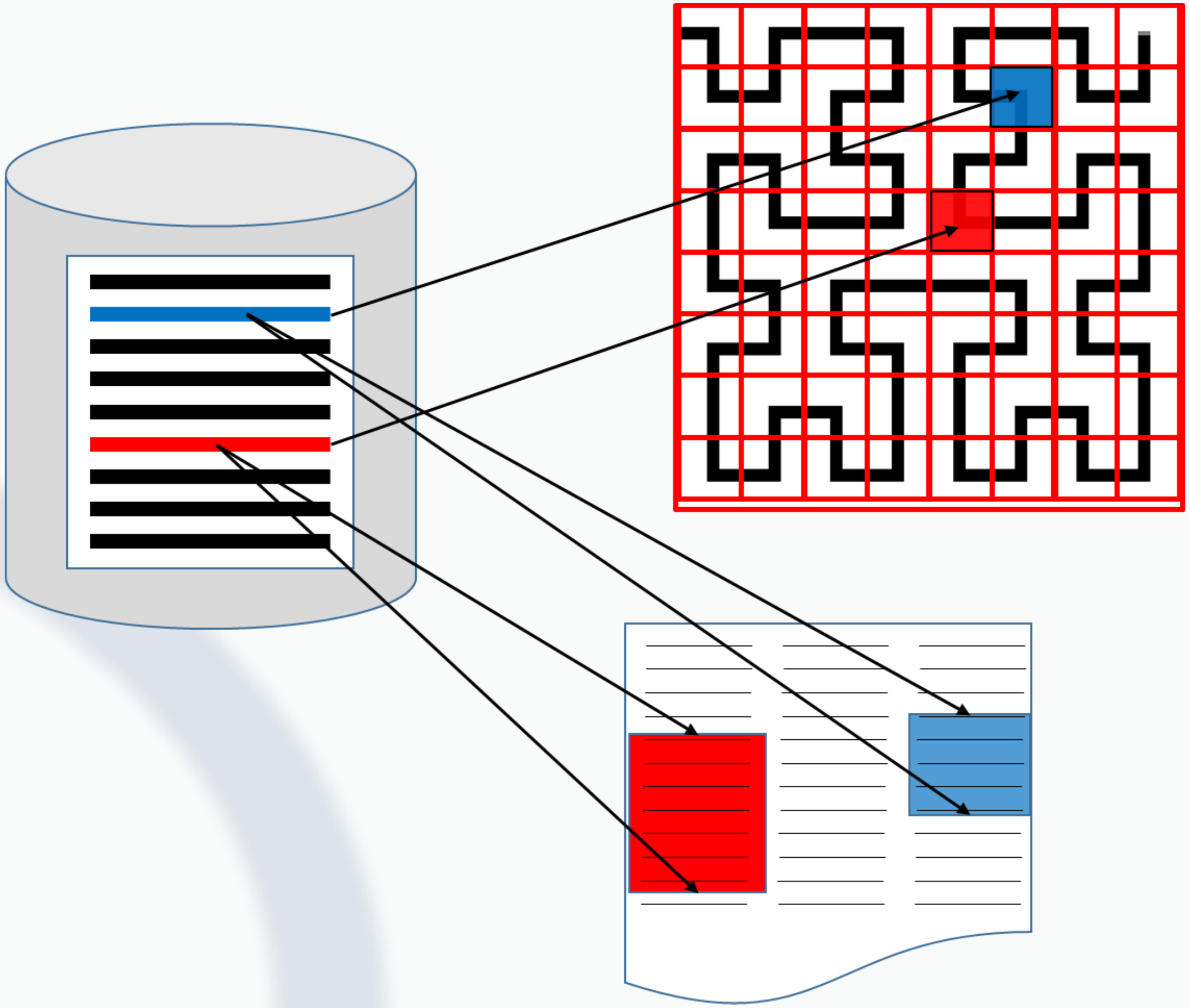


Individual files store intervals on curve; indexed in DB





DB also stores locations of “buckets” inside files.



SciServer

jupyter CosmoUC_5_PlotHalo Last Checkpoint: 11 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help

Dark-matter Halos in a Cosmological Simulation

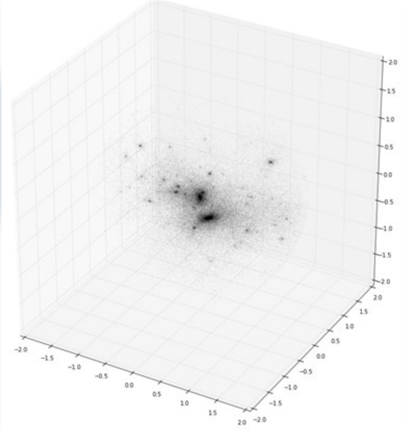
In [3]:
```
import SciServer.LoginPortal as Login
token = Login.getToken()
import SciServer.CasJobs
import pandas
import tables
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
```

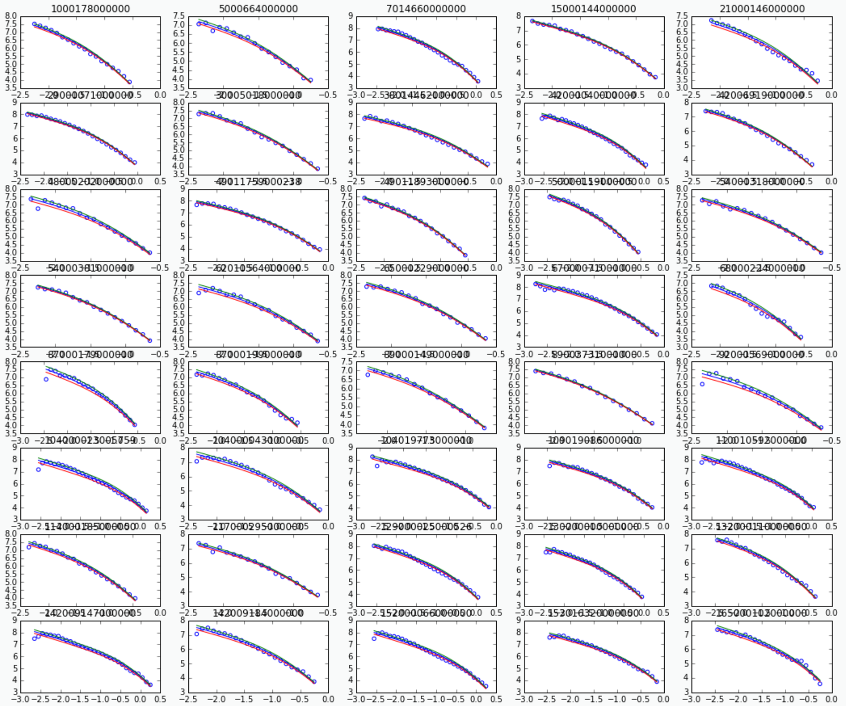
In [10]:
```
%%time
queryString = """
select top 100000 p.x-hh.x as x,p.y-hh.y as y,p.z-hh.z as z
from mpahalotrees.mr hh
    cross apply dbo.MillenniumParticles(hh.snapnum,
        dbo.Sphere::New(hh.x,hh.y,hh.z,3*hh.halfmassradius).ToString()) p
  where hh.haloid=84000007000000 order by newid()
      """
responseStream = SciServer.CasJobs.executeQuery(queryString, token=token,context="SimulationDB")
df = pandas.read_csv(responseStream, index_col=None)
```

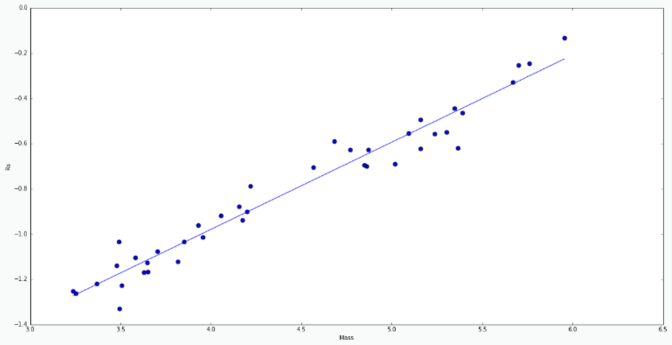
```
CPU times: user 351 ms, sys: 184 ms, total: 535 ms
Wall time: 5.27 s
```

In [13]:
```
fig = plt.figure(figsize=(15, 15))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.x,df.y, df.z,s=0.001)
```

Out[13]: <mpl_toolkits.mplot3d.art3d.Path3DCollection at 0x7fe86428b9e8>







Collab

jupyter ReadMillenniumSnapshots Last Checkpoint: 03/23/2017 (autosaved)

File Edit View Insert Cell Kernel Help

Python 3

```
    ax.set_title(title)
    divider = make_axes_locatable(ax)
    # Append axes to the right of ax3, with 20% width of ax3
    cax = divider.append_axes("right", size="5%", pad=0.05)
    # Create colorbar in the appended axes
    # Tick locations can be set with the kwarg `ticks`
    # and the format of the ticklabels with kwarg `format`
    cbar = plt.colorbar(im, cax=cax)#, ticks=MultipleLocator(0.2), format="%.2f")
```

```
CPU times: user 17.4 s, sys: 3.87 s, total: 21.3 s
Wall time: 27 s
```