



# From SkyServer to SciServer: the JHU DIBBs Project

19 August 2014



JOHNS HOPKINS  
UNIVERSITY

# Agenda

1. Introduction (Alex)
2. Background, Vision and Goals (Alex)
3. Science Collaboration (Alex)
4. SDSS Unification (Ani)
5. Project Operations, Roadmap and Progress (Mike)
6. Outreach and Collaboration (Jordan)
7. Summary (Alex)

# Big Data in Science

- ▶ Data growing exponentially, in all science
- ▶ All science is becoming data-driven
- ▶ This is happening very rapidly
- ▶ Data becoming increasingly open/public
- ▶ Non-incremental!
- ▶ Convergence of physical and life sciences through Big Data (statistics and computing)
- ▶ The “long tail” is important
- ▶ A scientific revolution in how discovery takes place

=> a rare and unique opportunity



# Science is Changing

## THOUSAND YEARS AGO

science was **empirical**  
describing natural phenomena



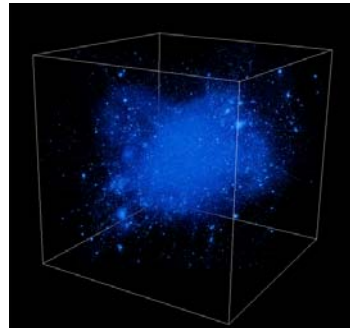
## LAST FEW HUNDRED YEARS

**theoretical** branch using models,  
generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$

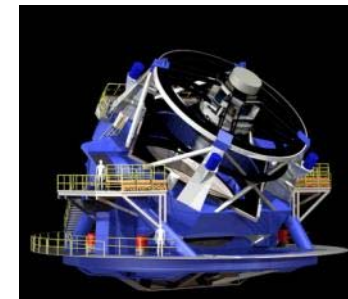
## LAST FEW DECADES

a **computational** branch simulating  
complex phenomena



## TODAY

**data intensive science**, synthesizing theory,  
experiment and computation with statistics  
►new way of thinking required!



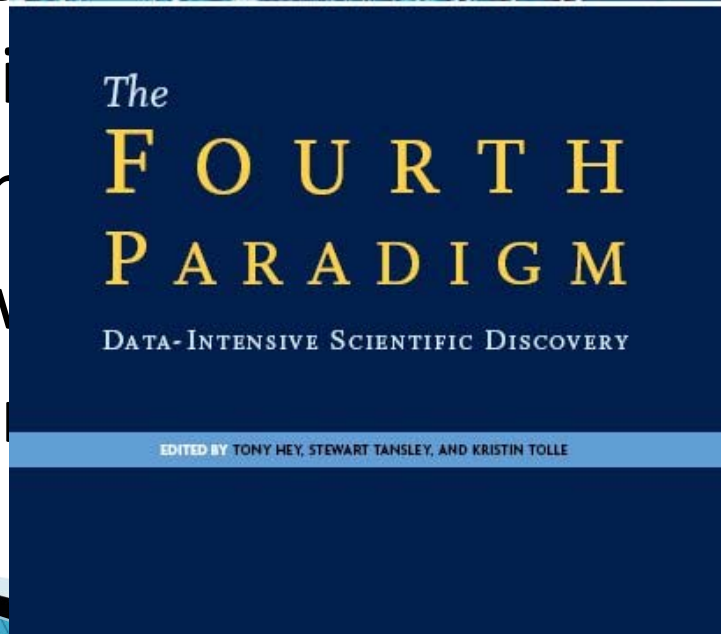
JOHNS HOPKINS  
UNIVERSITY



# Gray's Laws of Data Engineering

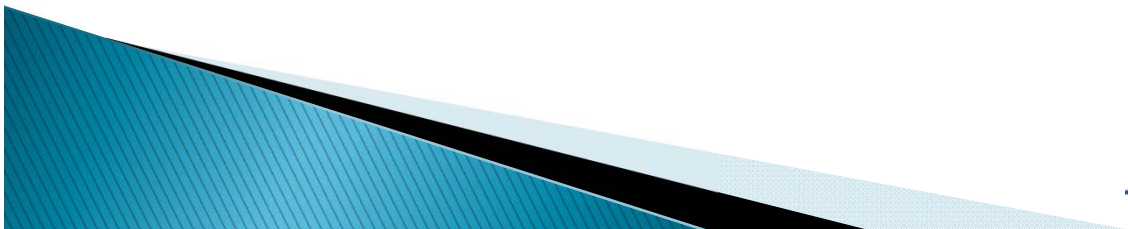
**Jim Gray**

- ▶ Scientific computing is moving around
- ▶ Need scientific analysis
- ▶ Take the data!
- ▶ Start with the data
- ▶ Go from data to knowledge



# Scientific Data Analysis Today

- ▶ Data grows as fast as our computing power
  - Consequence of Moore's Law and chip technology
  - Need tradeoff in analysis : best result in 1 min, 1 day, 1 month
  - Need randomized, incremental algorithms
  - Statistical vs systematic errors
- ▶ Need both “exploratory” and “confirmatory” searches
  - Quite different from “click-stream/ad-ware” data mining
- ▶ Data access patterns also quite different from business
  - Structured vs unstructured data
- ▶ Computer architectures CPU-heavy, IO-poor
- ▶ Universities hitting the “power wall”



# Exponential Data Growth

- ▶ How long does the data growth continue?
- ▶ High end always linear
- ▶ Exponential comes from technology + economics
  - ↔ rapidly changing generations
    - like CCD's replacing plates, and become ever cheaper
- ▶ How many new generations of instruments do we have left?
- ▶ Software is also an instrument
  - hierarchical data replication
  - virtual data
  - data cloning





# Data Access is Hitting a Wall

FTP and GREP are not adequate!

- ▶ You can GREP 1 MB in a second
- ▶ You can GREP 1 GB in a minute
- ▶ You can GREP 1 TB in 2 days
- ▶ You can GREP 1 PB in 3 years
- ▶ Oh!, and 1PB ~500 disks
- ▶ At some point you need **indices** to limit search  
**parallel** data search and analysis
- ▶ This is where **databases** can help
- ▶ You can FTP 1 MB in 1 sec
- ▶ You can FTP 1 GB / min (= 1\$/GB)
- ▶ ... 2 days and \$1K
- ▶ ... 3 years and \$1M



JOHNS HOPKINS  
UNIVERSITY

# Non-Incremental Changes

- ▶ Multi-faceted challenges
- ▶ New computational tools and strategies
- ▶ ... not just statistics, not just computer science, not just astronomy, not just genomics...
- ▶ Science is moving increasingly from hypothesis- driven to data-driven discoveries



# Why Is Astronomy Interesting?

Astronomy has always been data-driven....

Now becoming more accepted in other areas as well

- ▶ Important spatio-temporal features
- ▶ Very large density contrasts in populations
- ▶ Real errors and covariances
- ▶ Many signals very subtle, buried in systematics
- ▶ Data sets large, pushing scalability
  - LSST will be 100PB

*“Exciting, since it is **worthless!**”*

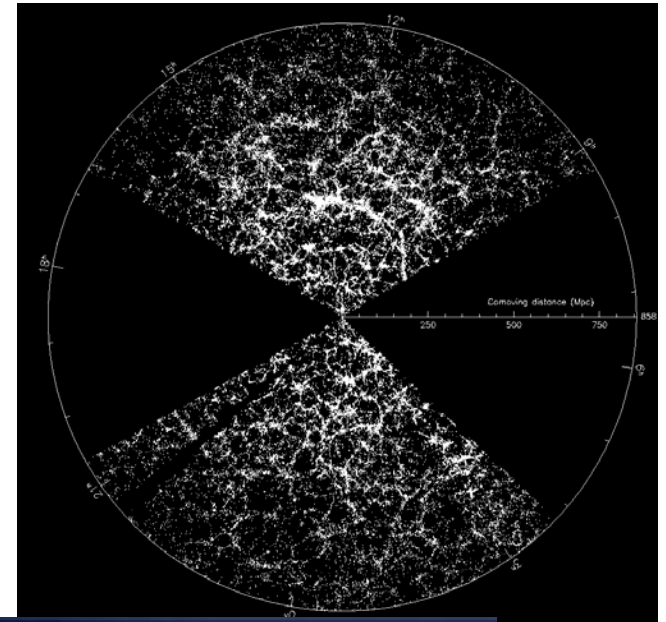
— Jim Gray





# Sloan Digital Sky Survey

- ▶ “The Cosmic Genome Project”
- ▶ Started in 1992, continuing through at least 2020
- ▶ Data is public
  - 2.5 Terapixels of images => 5 Tpx of sky
  - 10 TB of raw data => 100TB processed
  - 0.5 TB catalogs => 35TB in the end
- ▶ Database and spectrograph built at JHU (SkyServer)
- ▶ Data served from JHU



# Skyserver

- ▶ Prototype in 21st Century data access
  - 1.2B web hits in 12 years
  - 200M external SQL queries
  - 4,000,000 distinct users vs. 15,000 astronomers
  - The emergence of the “Internet Scientist”
  - The world’s most used astronomy facility today
  - Collaborative server-side analysis done by 7K astronomers



# Impact of Sky Surveys

## Astronomy

### Sloan Digital Sky Survey tops astronomy citation list

NASA's Sloan Digital Sky Survey (SDSS) is the most significant astronomical facility, according to an analysis of the 200 most cited papers in astronomy published in 2006. The survey, carried out by Juan Madrid from McMaster University in Canada and Duccio Macchetto from the Space Telescope Science Institute in Baltimore, puts NASA's Swift satellite in second place, with the Hubble Space Telescope in third (arXiv:0901.4552).

Madrid and Macchetto carried out their analysis by looking at the top 200 papers using NASA's Astrophysics Data System (ADS), which charts how many times each paper has been cited by other research papers. If a paper contains data taken only from one observatory or satellite, then that facility is awarded all the citations given to that article. However, if a paper is judged to contain data from different facilities – say half from SDSS and half from Swift – then both

#### Top 10 telescopes

Rank	Telescope	Citations	Ranking in 2004
1	Sloan Digital Sky Survey	1892	1
2	Swift	1523	N/A
3	Hubble Space Telescope	1078	3
4	European Southern Observatory	813	2
5	Keck	572	5
6	Canada–France–Hawaii Telescope	521	N/A
7	Spitzer	469	N/A
8	Chandra	381	7
9	Boomerang	376	N/A
10	High Energy Stereoscopic System	297	N/A

facilities are given 50% of the citations that paper received.

The researchers then totted up all the citations and produced a top 10 ranking (see table). Way out in front with 1892 citations is the SDSS, which has been

running since 2000 and uses the 2.5 m telescope at Apache Point in New Mexico to obtain images of more than a quarter of the sky. NASA's Swift satellite, which studies gamma-ray bursts, is second with 1523 citations, while the Hubble Space Telescope (1078 citations) is third.

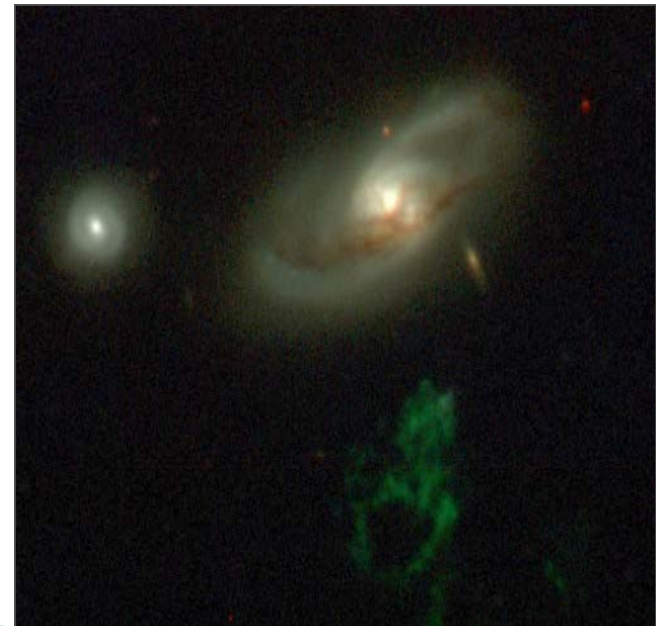
Although the 200 most cited papers make up only 0.2% of the references indexed by the ADS for papers published in 2006, those 200 papers account for 9.5% of the citations. Madrid and Macchetto also ignored theory papers on the basis that they do not directly use any telescope data. A similar study of papers published in 2004 also puts SDSS top with 1843 citations. This time, though, the European Southern Observatory, which has telescopes in Chile, comes second with 1365 citations and the Hubble Space Telescope takes third spot with 1124 citations.

Michael Banks

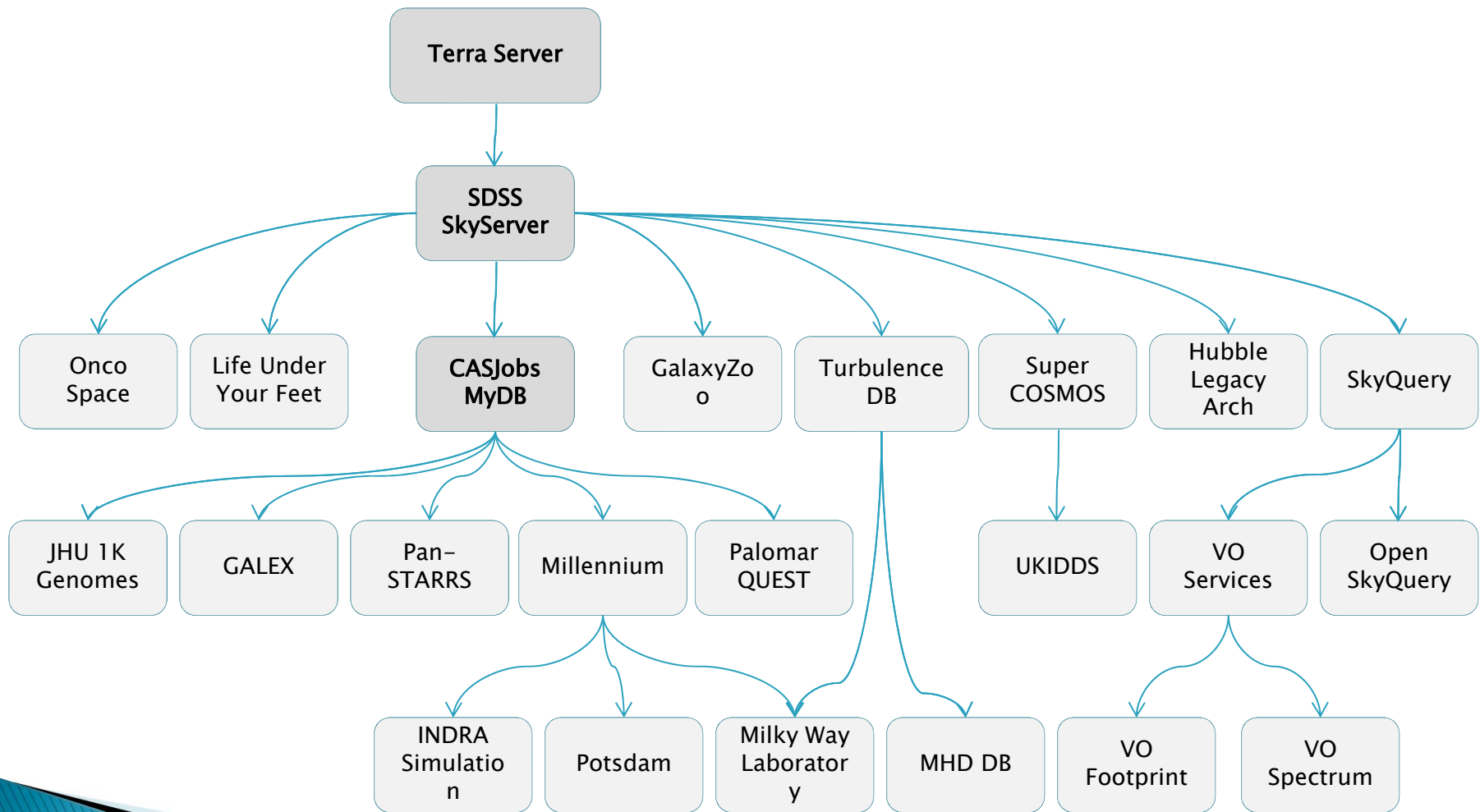


# GalaxyZoo

- ▶ 80 million visual galaxy classifications by the public
- ▶ Good publicity (CNN, Times, Washington Post, BBC)
- ▶ 500,000 people participating, blogs, poems...
- ▶ Original discoveries by the public (Voorwerp, Green Peas)
- ▶ Chris Lintott et al

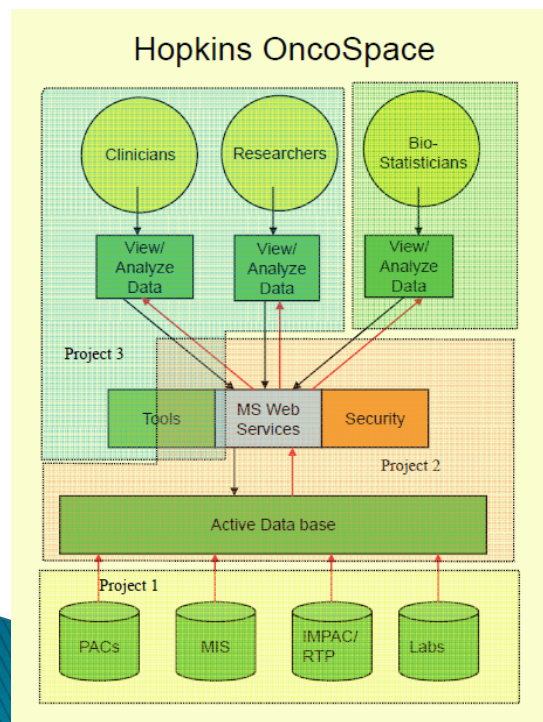


# The SDSS Genealogy



# Oncospace

- ▶ Todd McNutt, John Wong, JHU Radiation Oncology



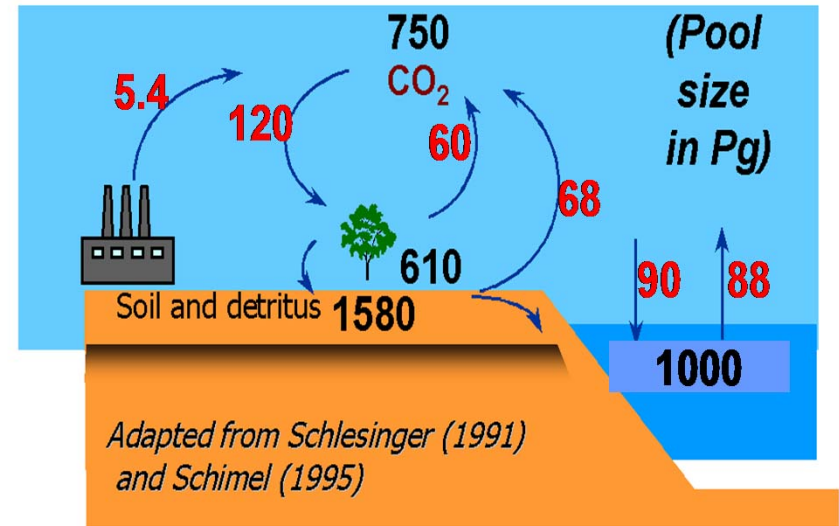
## OncoSpace: Adapting the SkyServer Approach

- **Active Databases**
- There is too much data to move around,  
***take the analysis to the data!***
- Do all data manipulations at database
  - ***Build custom procedures and functions in the database***
- Established Web-service for broad access
  - Query across multiple databases
- Automatic parallelism guaranteed

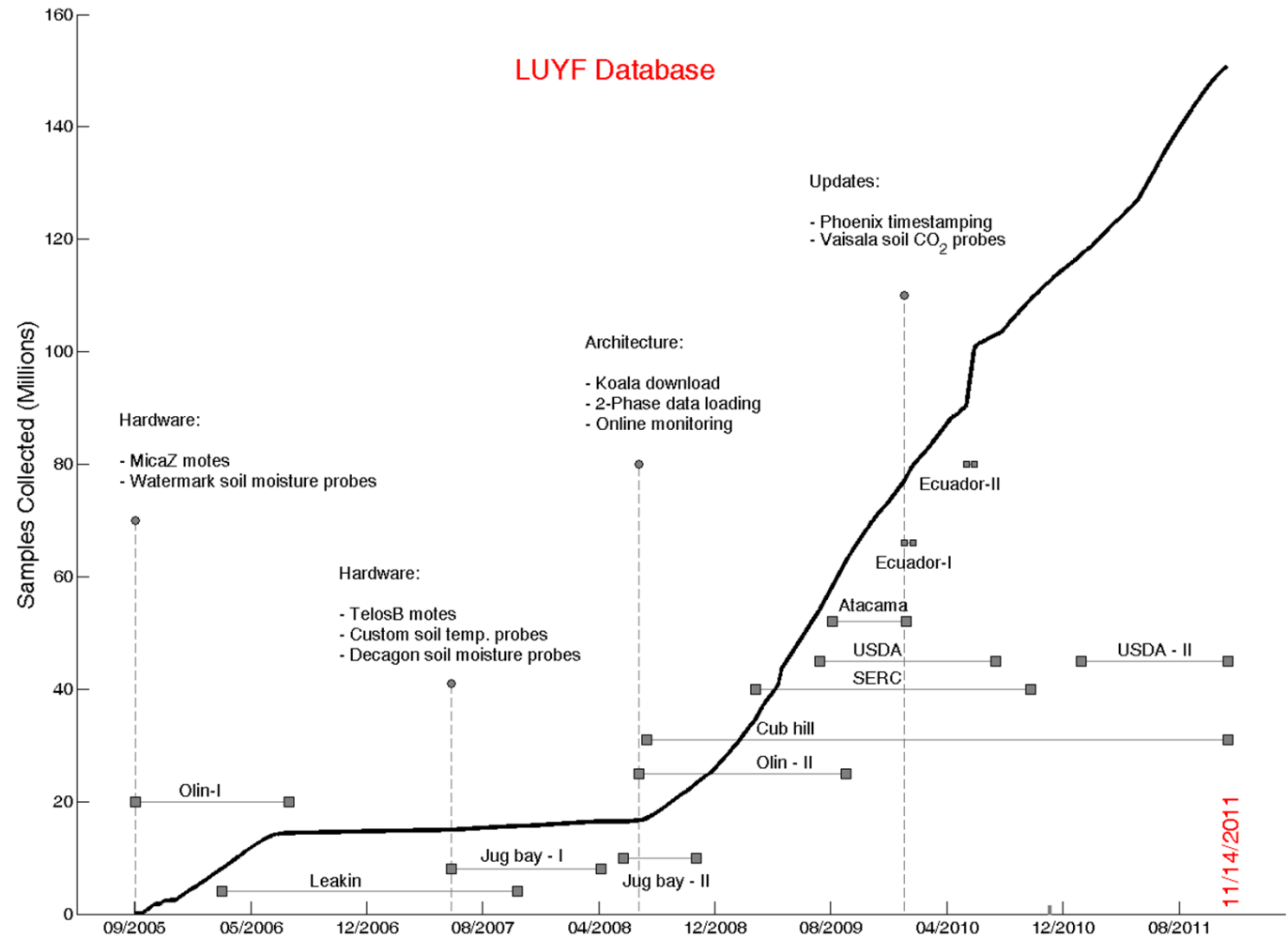


# Life Under Your Feet

- ▶ Role of the soil in Global Change
  - Soil CO<sub>2</sub> emission thought to be >15 times of anthropogenic
  - Using sensors we can measure it directly, in situ, over a large area
- ▶ Wireless sensor network
  - Use 100+ wireless computers (motest), with 10 sensors each, monitoring
    - Air +soil temperature, soil moisture, ...
    - Few sensors measure CO<sub>2</sub> concentration
  - Long-term continuous data, 180K sensor days, 30M samples
  - Complex database of sensor data, built from the SkyServer
  - End-to-end data system, with inventory and calibration databases
- ▶ with K.Szlavec (Earth and Planetary), A. Terzis (CS)
- ▶ <http://lifeunderyourfeet.org/>



# Cumulative Sensor Data



# Data in HPC Simulations

- ▶ HPC is an instrument in its own right
- ▶ Largest simulations approach petabytes
  - from supernovae to turbulence, biology and brain modeling
- ▶ Need public access to the best and latest through interactive numerical laboratories
- ▶ Creates new challenges in
  - How to move the petabytes of data (high speed networking)
  - How to look at it (render on top of the data, drive remotely)
  - How to interface (smart sensors, immersive analysis)
  - How to analyze (value added services, analytics, ... )
  - Architectures (supercomputers, DB servers, ??)

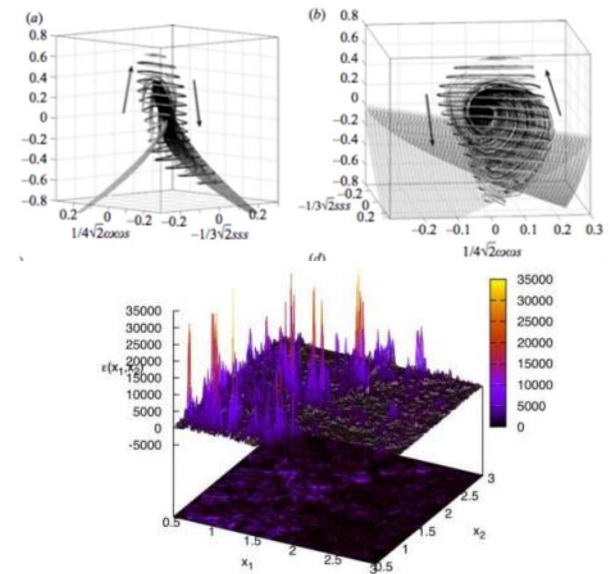


# Immersive Turbulence

*“... the last unsolved problem of classical physics...”*

## ► Understand the nature of turbulence

- Consecutive snapshots of a large simulation of turbulence:  
 $1024^4 \Rightarrow 30$  Terabytes
- Treat it as an experiment, **play** with the database!
- **Shoot test particles** (sensors) from your laptop into the simulation, like in the movie *Twister*



## ► New paradigm for analyzing simulations!

with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

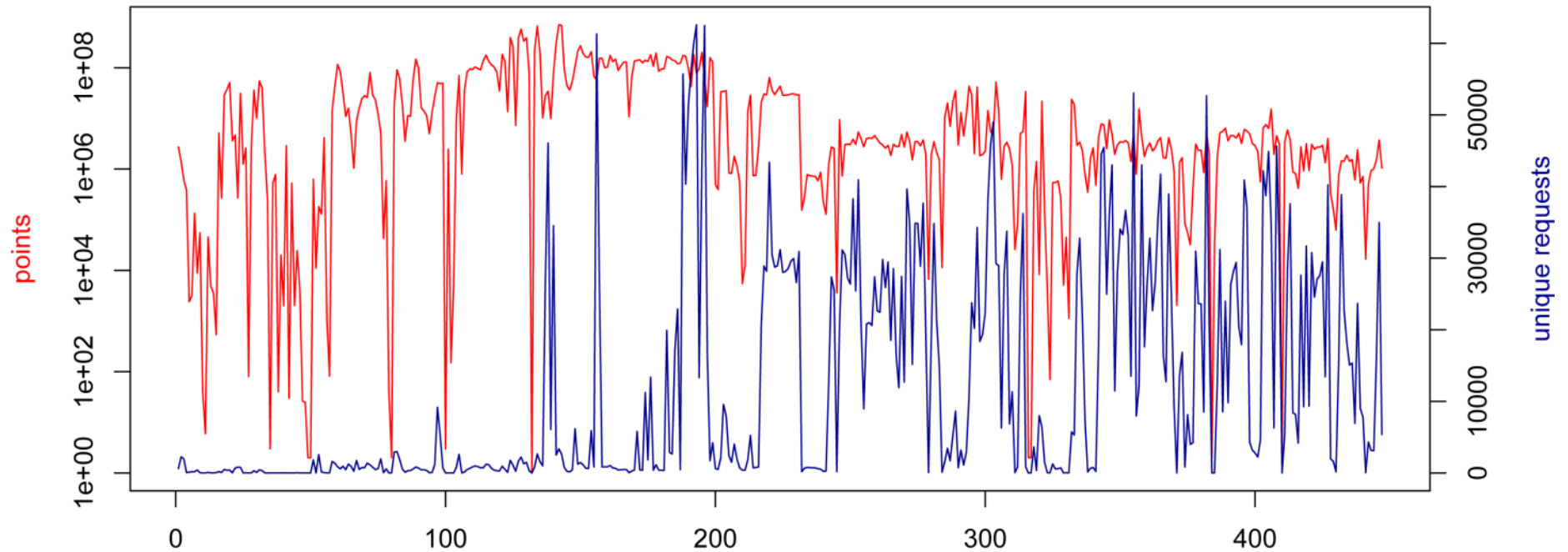


JOHNS HOPKINS  
UNIVERSITY



# Daily Usage

Turbulence Database Usage by Day



*Aug 2011: exceeded 100<sup>day</sup>B points, delivered publicly*

*Aug 2014: exceeded 7.5T points*

# Simulations in the DB

- ▶ Amazing progress in 7 years
- ▶ Millennium and turbulence are showcases
- ▶ People playing the DB like a musical instrument
- ▶ New challenges emerging:
  - Petabytes of data, trillions of particles
  - Increasingly sophisticated value added services
  - Need a coherent strategy to go to the next level
- ▶ Not just storage, but integrate access and computation
- ▶ Filling the gap between DB server and supercomputer
- ▶ The democratization of supercomputers



# Scalable Data-Intensive Analysis

- ▶ Large data sets => data resides on hard disks
- ▶ Analysis has to move to the data
- ▶ Hard disks are becoming sequential devices
  - For a PB data set you cannot use a random access pattern
- ▶ Both analysis and visualization become streaming problems
- ▶ Same thing is true with searches
  - Massively parallel sequential crawlers (MR, Hadoop, etc)
- ▶ Spatial indexing needs to be maximally sequential
  - Space filling curves (Peano-Hilbert, Morton,...)



# The Long Tail

- ▶ The “Long Tail” of a huge number of small data sets
  - The integral of the “long tail” is big!
- ▶ Facebook: bring many small, seemingly unrelated data to a single place and new value emerges
  - What is the science equivalent?
- ▶ The DropBox lesson
  - Simple interfaces are more powerful than complex ones
  - Interface is open, public
- ▶ SciDrive: JHU project (funded by the Sloan Foundation)
  - Enable people to drag and drop (and share) their data
  - No metadata required
  - We are trying to figure it out from the data itself + papers



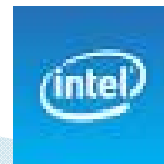


# JHU Data-Scope

- ▶ Funded by NSF MRI to build a new ‘instrument’ to look at data
- ▶ Goal: ~100 servers for \$1M + about \$200K switches+racks
- ▶ Two-tier: performance (P) and storage (S)
- ▶ Large (6.5PB) + cheap + fast (500GBps), but ...
  - ..a special purpose instrument
- ▶ 100G connectivity to the outside world



	O					
	1P	1S	All P	All S	Full	
servers	1	1	90	6	102	
rack units	4	34	360	204	564	
capacity	24	720	2160	4320	6480	TB
price	8.8	57	8.8	57	792	\$K
power	1.4	10	126	60	186	kW
GPU*	1.35	0	121.5	0	122	TF
seq IO	5.3	3.8	477	23	500	GBps
IOPS	240	54	21600	324	21924	kIOPS
netwk bw	10	20	900	240	1140	Gbps



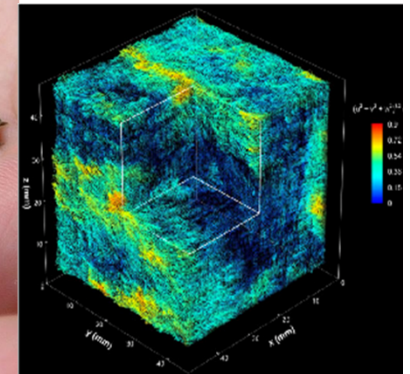
# SciServer

- ▶ Build on the 20 years of Astronomy experience
- ▶ Reengineer the SkyServer into a broad platform
- ▶ Deal with the “service lifecycle”
- ▶ Systematically engage a much broader spectrum
- ▶ Focus: database-centric science at scale

*“Do not try to be everything for everybody”*

# Collaborative Science Projects

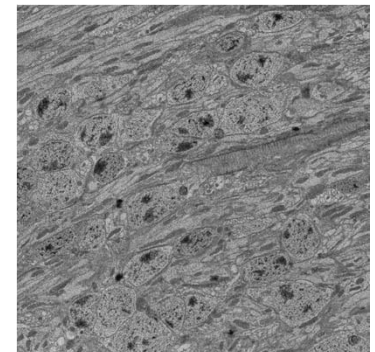
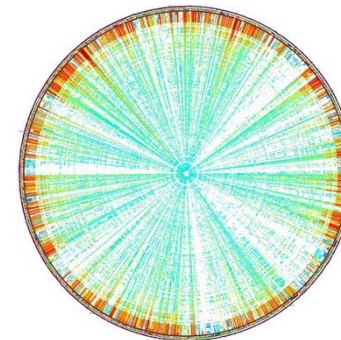
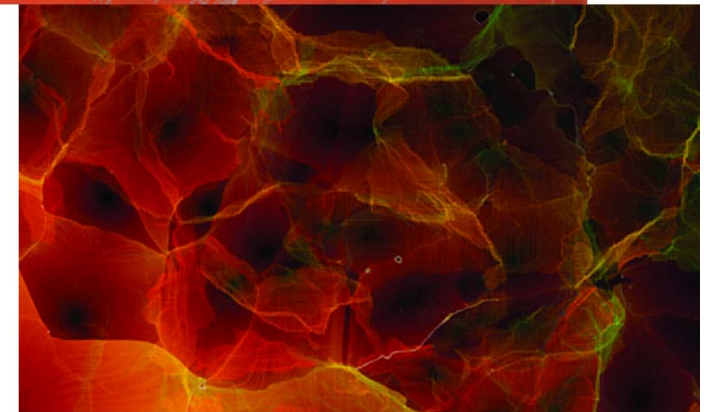
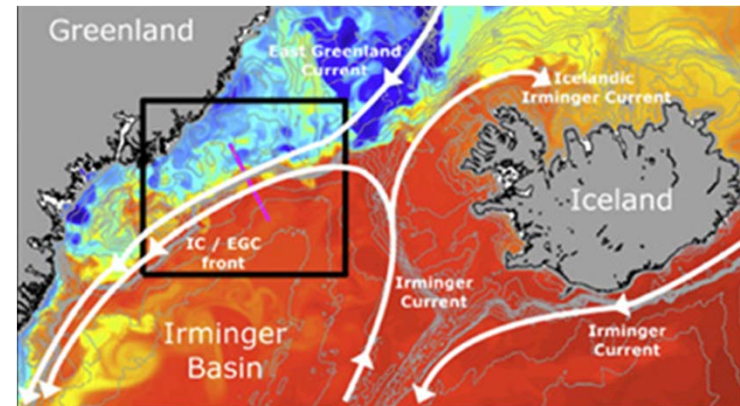
- ▶ **Astronomy** (Alex Szalay): the SDSS map of the Universe, extending to larger datasets
- ▶ **Soil Ecology** (Katalin Szlavecz): wireless sensors for long-term environmental monitoring by researchers and citizen scientists
- ▶ **Turbulence** (Charles Meneveau): large-scale numerical simulations of turbulent flow





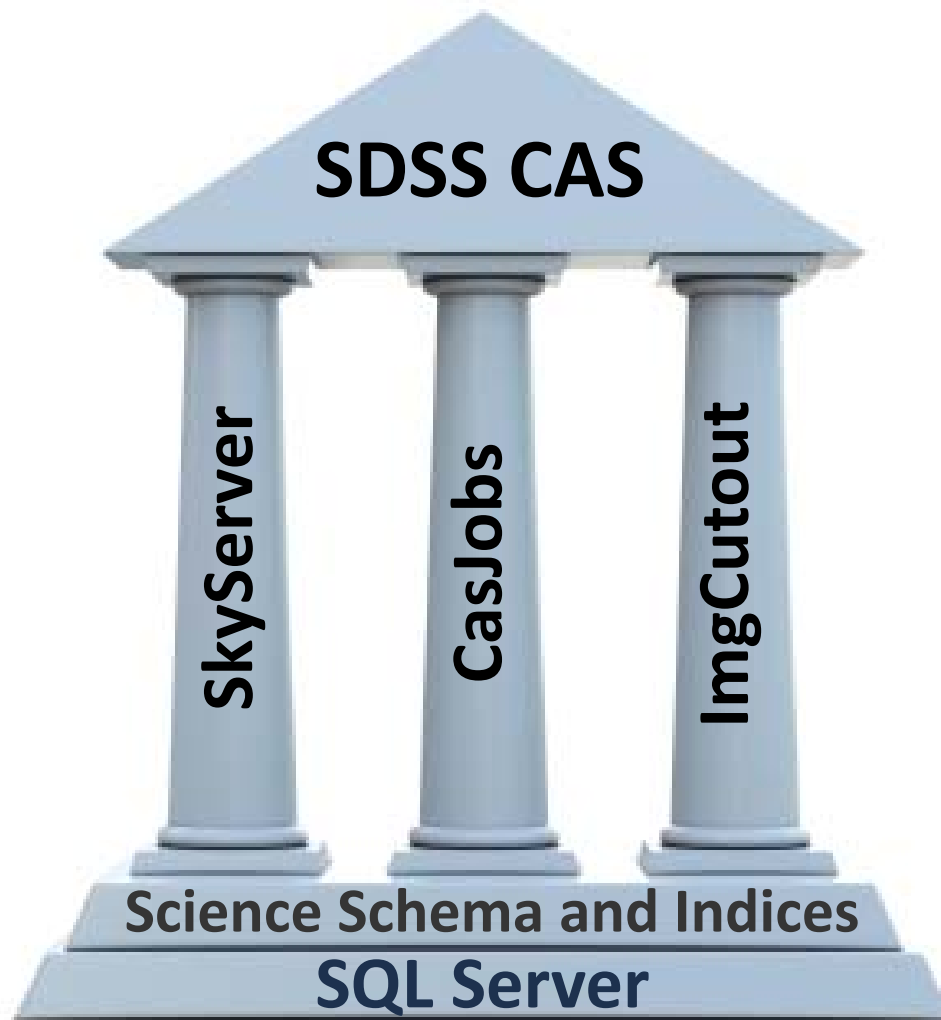
# Collaborative Science Projects

- ▶ **Oceanography** (Tom Haine): models of the km-scale spatiotemporal evolution of ocean currents
- ▶ **Cosmological Physics** (Gerard Lemson): simulations of the evolution of galaxies in the Universe
- ▶ **Genomics** (Steven Salzberg): cross-correlation of entire organism genomes
- ▶ **Connectomics** (Randal Burns): mapping connections in neural networks





# The SDSS “CAStle”

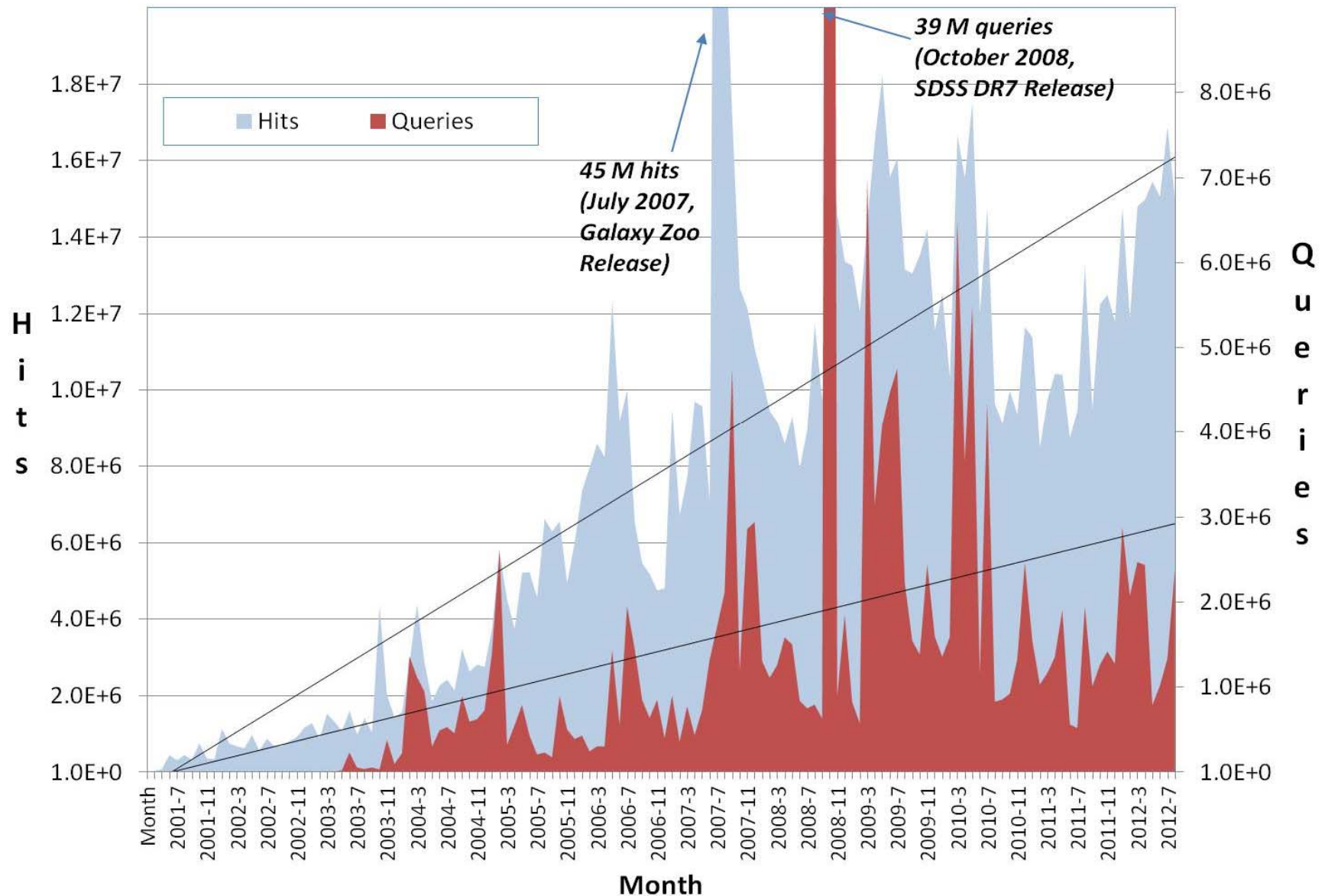


- ▶ 3 pillars
  - Synchronous access for quick queries
  - Asynchronous access for long, intensive queries
  - Visual browsing of images
- ▶ Everything served from commercial RDBMS
- ▶ Extensively indexed
- ▶ Science built into schema

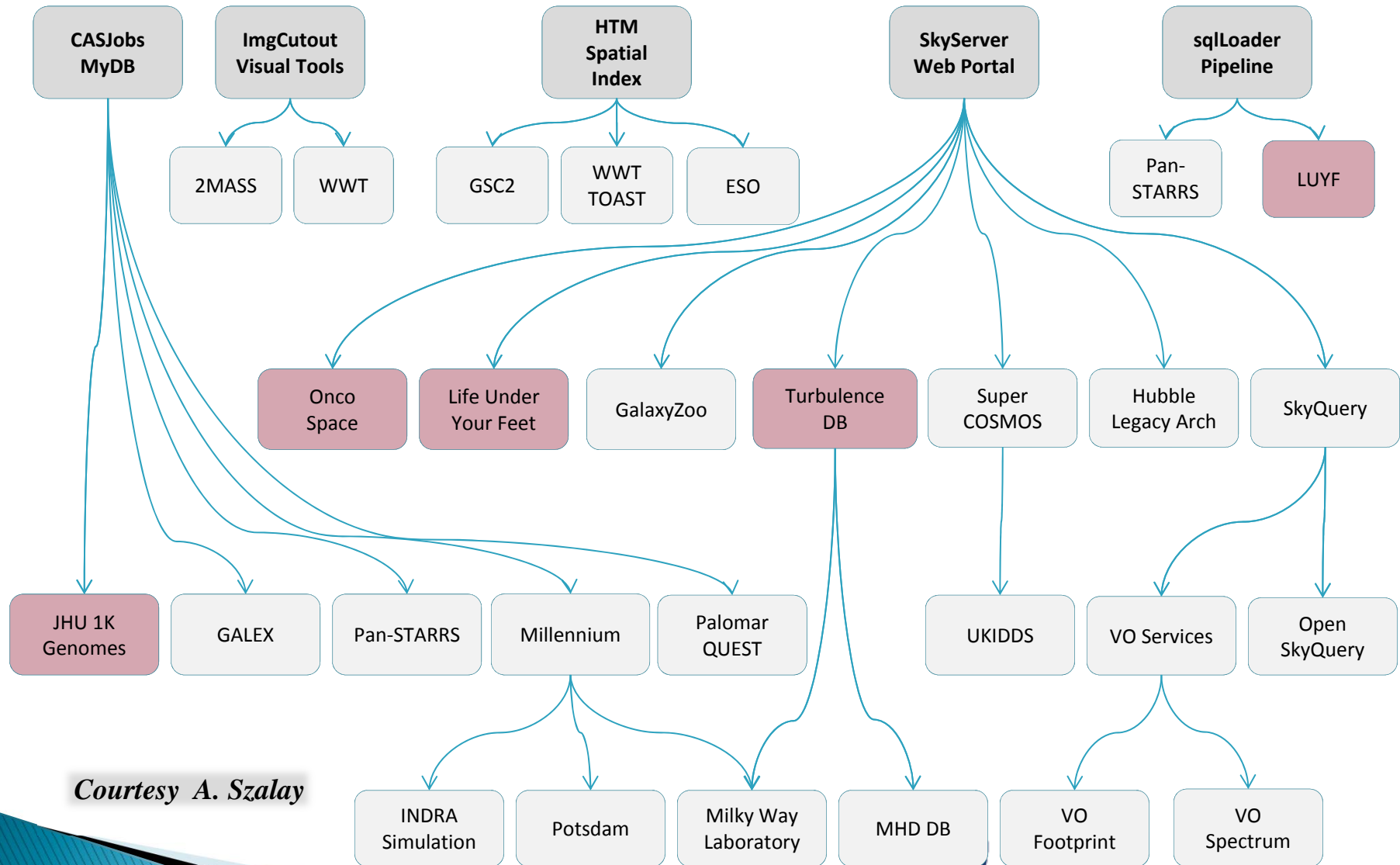


# How Skyserver changed Astronomy

## Monthly Web Hits and SQL Queries



# The SDSS Genealogy (updated)



*Courtesy A. Szalay*

# The Broad Impact of SDSS

- ▶ 5,000 publications, 200,000 citations
  - More papers from *outside* the collaboration
- ▶ Changed the way we do astronomy
- ▶ Remarkably fast transition seen for the community
- ▶ Speeded up the first phase of exploration
- ▶ Wide-area statistical queries easy
- ▶ Multi-wavelength astronomy is the norm
- ▶ SDSS earned the TRUST of the community
- ▶ Enormous number of projects, way beyond original vision and expectation
- ▶ Many other surveys now follow
- ▶ Established expectations for data delivery
- ▶ Serves as a model for other communities of science



# Reusable Building Blocks

- ▶ SkyServer
  - Extensive built-in science, query, metadata support
- ▶ CasJobs batch query workbench
  - Collaborative science in the big data era
  - Adapted and deployed in several (non) astro projects
- ▶ ImgCutout visual image browsing service
  - Recently adapted to display non-SDSS images
- ▶ Hierarchical Triangular Mesh spatial index
  - CLR library written in C#, ported to Java, C++
- ▶ sqlLoader data loading pipeline
  - Automates the task of ingesting Terabytes of data into DBs

# Reusable Building Blocks

- ▶ GrayWulf: scalable software/hardware for Big Data
  - The “DataWulf” concept renamed in honor of Jim Gray, who pioneered it
  - “Bring the program to the data”, not vice-versa
- ▶ SkyQuery: federated cross-matching service
  - Built on top of GrayWulf
  - Enables multi-wavelength astronomy
  - The holy grail of the Virtual Observatory
- ▶ SciDrive: data hosting/sharing for VO community
  - DropBox-like service based on VOSpace standard
  - Built upon OpenStack/Swift
  - Solution for long-tail astronomy – instantly connects user data with VO data universe

# SkyServer

Object Explorer

skyserver.sdss3.org/dr10/en/tools/explore/summary.aspx?id=0x112d14c220880060&spec=0x28e84d919a006...

**DR10**

Explore Home

Search

Imaging Summary

FITS

Finding chart

Other Observations

Neighbors

Galaxy Zoo

PhotoTag

Field

PhotoObj

PhotoZ

PhotoZRF

Cross-ID

Spec Summary

FITS

Plate

All Spectra

SpecObj

sppLines

galSpecLine

galSpecIndx

galSpecInfo

Fit Parameters

sppParams

StarformingPort

PassivePort

emissionLinesPort

PCAWscBC03

PCAWscM11

FSPSGranEarlyDust

FSPSGranEarlyNoDust

FSPSGranWideDust

FSPSGranWideNoDust

IR Spec Summary

ApogeeStar

SDSS

**SDSS J131027.46+182617.4**

Look up common name

Type		SDSS Object ID	
GALAXY		1237668296598749280	
RA, Dec		Galactic Coordinates (l, b)	
Decimal	Sexagesimal	l	b
197.61446, 18.43817	13:10:27.46, +18:26:17.40	330.66079	80.26964

**Imaging WARNING:** This object's photometry may be unreliable. See the photometric flags below.

Flags: DEBLEND\_DEGENERATE PSF\_FLUX\_INTERP DEBLENDED\_AT\_EDGE BAD\_MOVING\_FIT BINNED1 INTERP COSMIC\_RAY NODEBLEND CHILD BLENDED

**Magnitudes**

u	g	r	i	z
16.52	14.95	14.00	13.32	13.32

**Magnitude uncertainties**

err_u	err_g	err_r	err_i	err_z
0.01	0.00	0.00	0.02	0.00

**Image**

Image MJD	mode	Other observations	parentID	nChild	extinction_r	PetroR <sub>a</sub> (arcmin)
53500	PRIMARY	0	1237668296598749279	0	0.06	18.23 ± 2.0

photoZ (KD-tree method) 0.154 ± 0.0550 photoZ (RF method) 0.164 ± 0.1785 Galaxy Zoo 1 morphology Uncertain

**Cross-identifications** [Show](#)

**Optical Spectra** SpecObjID= 2947691243863304192 [Interactive spectrum](#)

Survey: sdss Progress: legacy Target: GALAXY\_102 GALAXY\_102  
RA: 13:10:27.46, Dec: 18:26:17.40, Filter: SDSS\_r, Filter: SDSS\_r, SDSS\_r  
objID: 0x112d14c220880060, Class: GALAXY, SDSS\_r

**Spectrograph**

class	Redshift (z)	Redshift err
GALAXY	0.012	0.00001

skyserver.sdss3.org/dr10/en/tools/search/sql.aspx

**SLOAN DIGITAL SKY SURVEY III**

**SkyServer DR10**

Home Data Schema Education Astronomy SDSS Contact Us Download Site Search Help

**DR10 Tools**

Getting Started

Famous places

Get Images

Scrolling sky

Visual Tools

Search

- Radial
- Rectangular
- Search Form
- SQL
- Imaging Query
- Spectro Query
- IR Spec Query

Object Crossid

CasJobs

**SQL Search**

This page allows you to directly submit a SQL (Structured Query Language) query to the SDSS database server. You can modify the default query as you wish, or cut and paste a query from the [SDSS Sample Queries](#) page.

**Please note:** To be fair to other users, queries run from SkyServer search tools are restricted in how long they can run and how much output they return, by **timeouts** and **row limits**. Please see the [Query Limits help](#) page. To run a query that is not restricted by a timeout or number of rows returned, please use the [CasJobs batch query service](#).

Clear Query

-- This query does a table JOIN between the imaging (PhotoObj) and spectra  
-- (SpecObj) tables and includes the necessary columns in the SELECT to upload  
-- the results to the SAS (Science Archive Server) for FITS file retrieval.

```
SELECT TOP 10
  p.objid, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z,
  p.run, p.rerun, p.camcol, p.field,
  s.specobjid, s.class, s.z, s.z_err,
  s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  p.u BETWEEN 0 AND 19.6
  AND g BETWEEN 0 AND 20
```

Submit ☐ Check Syntax Only? **Output Format** HTML XML CSV JSON VOTable FITS Reset

skyserver.sdss3.org/dr10/en/tools/chart/list.aspx

**DR10**

Home Help Chart Nav Explore

Use query to fill form

name,ra,dec  
1466095173023852544,141.5  
2657265679694063616,141.5  
534863786490750976,141.95  
2184269790409943040,141.5  
533706001333905408,141.95

Cut and paste ra/dec list

**Parameters**

scale 0.4 "/pix

opt

**Get Image**

**Drawing options**

- ☐ Grid
- ☐ Label
- ☐ Photometric objects
- ☐ Objects with spectra
- ☐ Invert Image

**Advanced options**

- ☐ APOGEE Spectra
- ☐ SDSS Outlines
- ☐ SDSS Bounding Boxes
- ☐ SDSS Fields
- ☐ SDSS Masks
- ☐ SDSS Plates

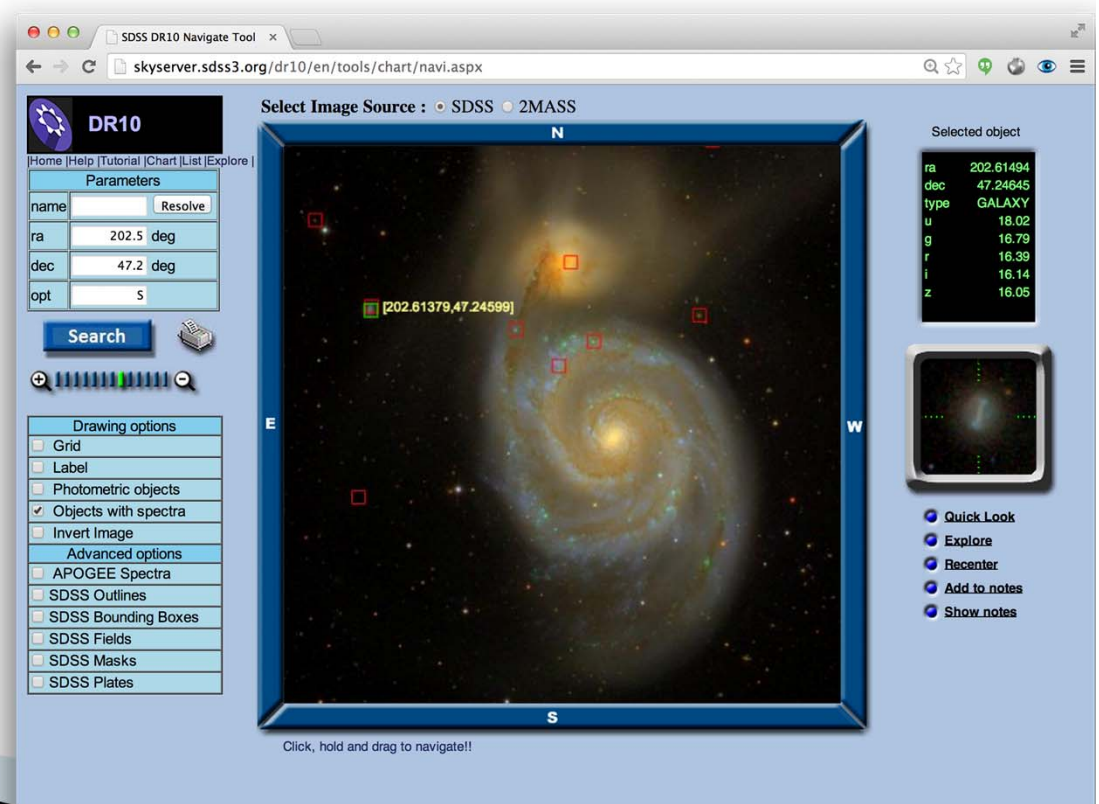
obj list page 1 page 2

1466095173023852544 J092759.09+085342.4	2657265679694063616 J092759.22+184202.8	534863786490750976 J092759.25+013821.8	2184269790409943040 J092759.45+261928.2	533706001333905408 J092759.7-002248.6
258180812085699776 J092800.233229.6	1434564481966434304 J092800.05+363217.4	2183241747138635776 J092800.34+304533.4	2184372319869233152 J092800.39+275310.5	1434566131233875968 J092800.58+364012.4
534865710636095584 J092800.63+014017.4	218324022016542720 J092800.92+300213	2111480501170825216 J092801.1+661518.8	2185432249128740864 J092801.19+301154.1	2185422903270904768 J092801.29+312214.9
639662917507115008 J092801.58+040041.8	86368951144548352 J092801.6+524052.9	1346660172197554176 J092801.68+070737.6	1793678448548407296 J092801.72+342437.9	534872857461680128 J092801.82+005820.4
218548502668674112 J092801.83+321334.1	2184375343526208536 J092802.04+271531.9	2185430874739206144 J092802.17+304536.7	2184271439677384704 J092802.17+261050.4	5226442585270648832 J092802.47+372636.1

- ▶ Extensive built-in science, query, metadata support

# ImgCutout

- ▶ Visual image browsing service
  - Recently adapted to display non-SDSS images
- ▶ Uses Hierarchical Triangular Mesh spatial index
  - CLR library written in C#, ported to Java, C++





# CAS Jobs

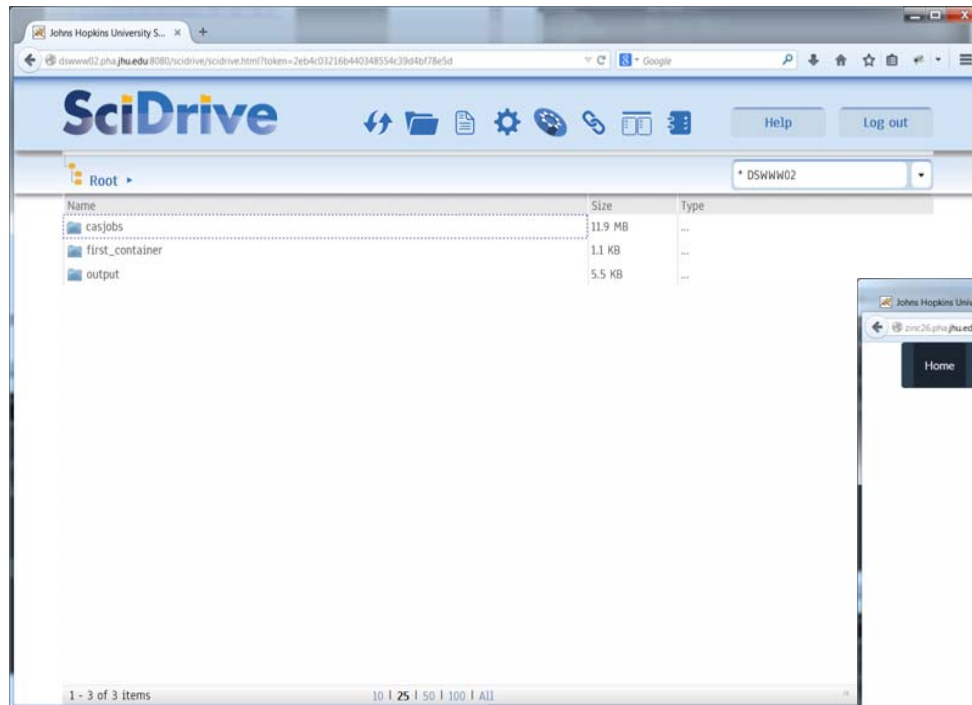
- ▶ Batch query workbench
- ▶ Adapted and deployed in several (non) astronomy projects

The screenshot displays the SDSS Query / CasJobs web interface. The browser address bar shows 'skyserver.sdss3.org/CasJobs/SubmitJob.aspx'. The interface includes a navigation menu with options like Help, Tools, Query, History, MyDB, Import, Groups, Output, Profile, Queues, SkyServer, Logout, and a user name 'deeppm'. Below the menu, there's a 'Context' section with 'Table (optional)' set to 'MyTable' and 'Task Name' set to 'My Query'. A SQL query is entered in the 'Samples' field:   

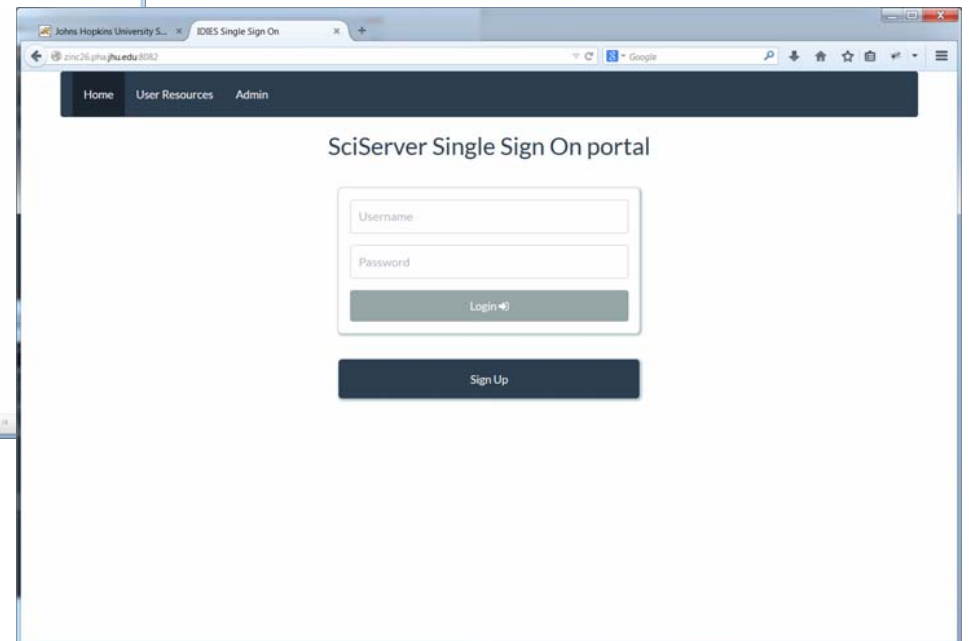
```
1 select top 40 p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z
2 from photoobj p
3 join dbo.fgetnearbyobjeq(44.41, 5.99, 40) n on p.objid=n.objid
4 where p.g between 14 and 18
```

The query status is 'Query complete!'. Below the query, there are buttons for 'Syntax', 'Plan', 'Quick', and 'Submit'. The results are displayed as a table with 40 rows and columns: objid, ra, dec, u, g, r, i, z. The table contains numerical data for each row. At the bottom, there's a 'Notes' section titled 'ConeSearchClient' with a date '1/1/2001 12:00:00 AM'. Below this is a 'PMStar' section with a description: 'Proper motion of a star (point source) in the Deep PM catalog.' It explains that each SDSS DR7 star that falls within the survey area, and within a given r magnitude range, is listed in this table. It also mentions that the number of matching objects in SDSS is indicated by a digit (0-10) and that 11 indicates a one-to-one match. If 0, there was no match for this SDSS object in our survey, and all remaining columns for this object will be set to 0.

# SciDrive & Portal



- ▶ “Dropbox-like” for easy data upload and sharing
  - Built on OpenStack/Swift



# SkyQuery

**MyDB Summary**

The following is a summary of your MyDB usage.

MyDB Usage:

Space allocated: 4 MB

Space used: 2.195 MB

Log space allocated: 1 MB

• Running out of space? You can request more.

**Download tables**

Download tables of MyDB into various data file formats.

- All exported files are automatically compressed with zip compression.
- File names are generated from table names automatically.
- Use batch export for large tables.

Tables:

- dbo.archive\_tar\_txt\_numbers
- dbo.newtable
- dbo.TestData
- dbo.txt\_numbers\_txt\_numbers

File format:

- TXT (tabulated text file)**
- CSV (delimited text file)
- XHTML file
- SQL Server native format

Job ID	Type	Submitted	Started	Finished	Status	Comments
test_14081515571750227	Import	8/15/2014 3:57:17 PM			waiting	test comments
test_14081515565074649	Import	8/15/2014 3:56:50 PM			waiting	test comments
test_14081515564221832	Import	8/15/2014 3:56:42 PM			waiting	test comments
test_14081515434695152	Export	8/15/2014 3:43:46 PM			waiting	test comments
test_1408151542754744	Export	8/15/2014 3:14:27 PM			waiting	test comments
test_14081515104621983	Export	8/15/2014 3:10:46 PM			waiting	test comments
test_14081121453971427	Query	8/11/2014 9:45:39 PM	8/11/2014 9:45:44 PM	8/11/2014 9:45:45 PM	failed	testjob
test_14081121424123202	Export	8/11/2014 9:42:41 PM	8/11/2014 9:42:46 PM	8/11/2014 9:42:46 PM	completed	
test_14081121415803753	Query	8/11/2014 9:41:58 PM	8/11/2014 9:42:02 PM	8/11/2014 9:42:27 PM	failed	testjob
test_14081121414768477	Import	8/11/2014 9:41:47 PM	8/11/2014 9:41:52 PM	8/11/2014 9:41:52 PM	completed	comments
test_14081121413706682	Import	8/11/2014 9:41:36 PM	8/11/2014 9:41:42 PM	8/11/2014 9:41:42 PM	completed	comments
test_14081121402949350	Import	8/11/2014 9:40:29 PM			canceled	comments
test_14081121393370331	Import	8/11/2014 9:39:33 PM			canceled	comments
test_14081121391614316	Query	8/11/2014 9:39:16 PM	8/11/2014 9:40:30 PM		canceled	testjob
test_14081121385049846	Import	8/11/2014 9:38:50 PM	8/11/2014 9:39:21 PM	8/11/2014 9:39:33 PM	canceled	comments
test_14081121364612147	Export	8/11/2014 9:36:45 PM	8/11/2014 9:39:15 PM	8/11/2014 9:39:16 PM	canceled	
test_14081121354830650	Query	8/11/2014 9:35:48 PM	8/11/2014 9:39:04 PM	8/11/2014 9:39:15 PM	canceled	testjob

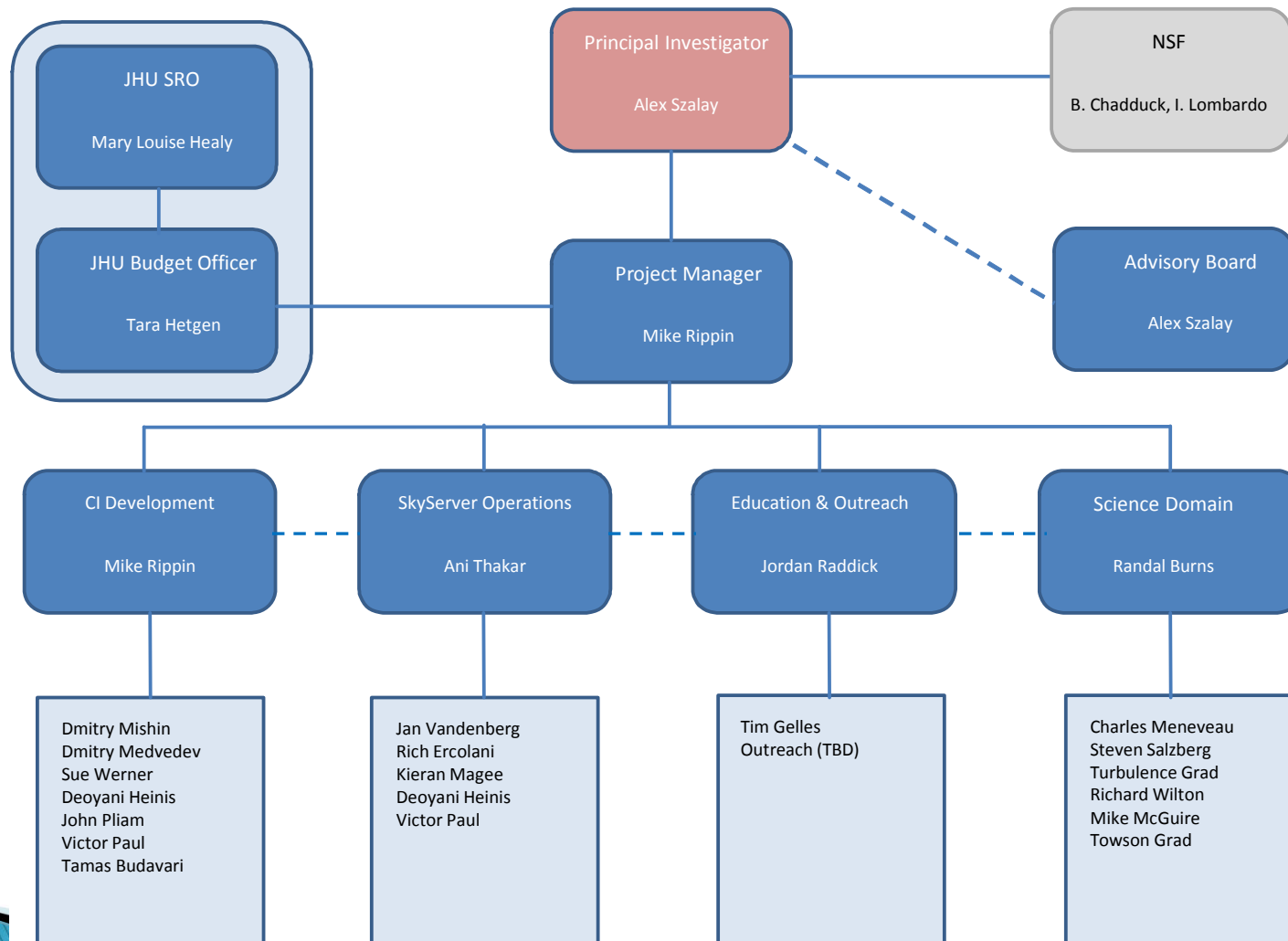
- ▶ Enables multi-wavelength astronomy
- ▶ The holy grail of the Virtual Observatory

# Project Management

- ▶ NSF Cooperative Agreement
- ▶ Project Execution Plan – 5 years
- ▶ Yearly assessment + 18 and 36 Month Reviews
- ▶ Structured around four Core Functions
  - CI Development
  - SDSS Unification
  - Science Domain Collaboration
  - Outreach and Education
- ▶ All work together
- ▶ Science Driven
- ▶ User groups
- ▶ External Advisory Board



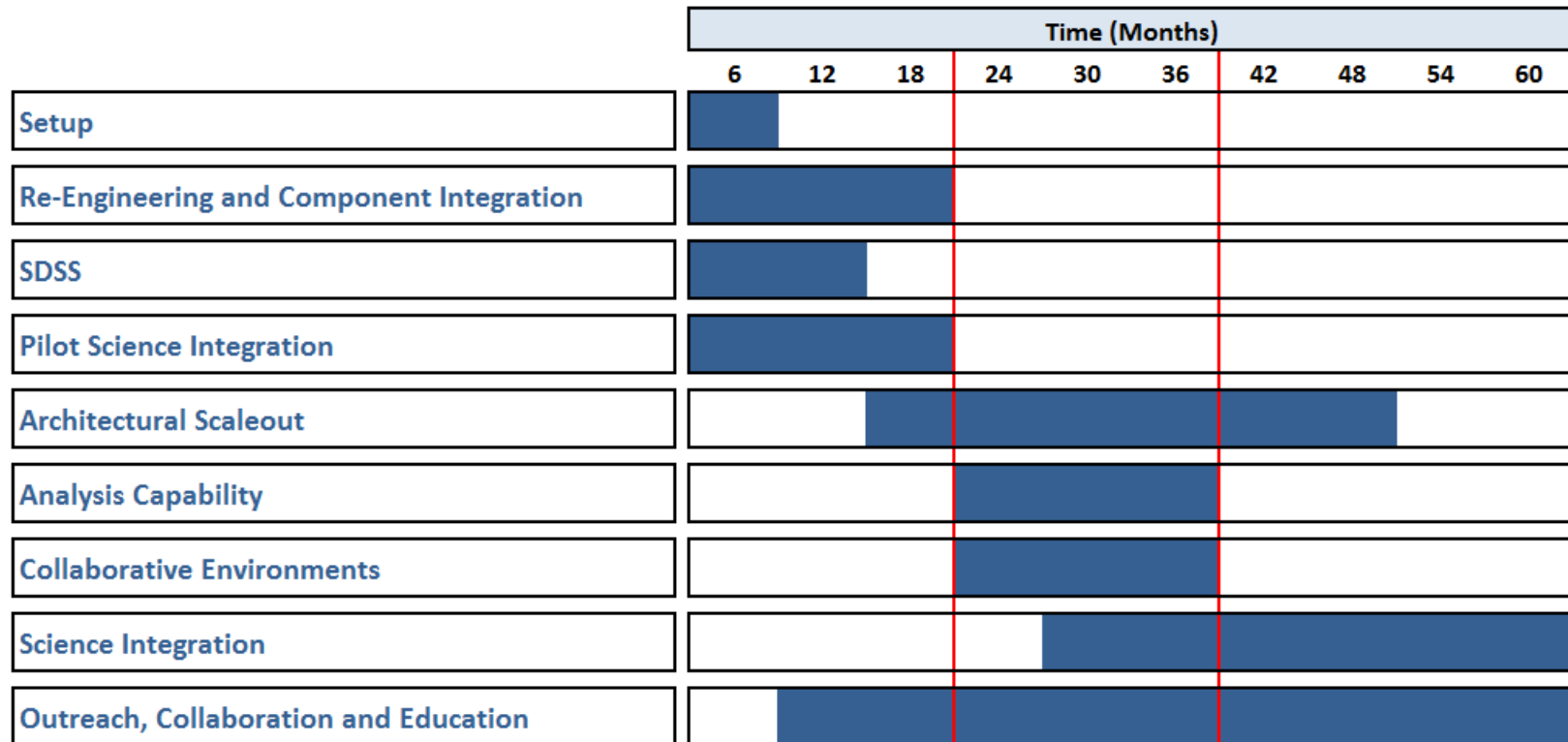
# Teams and Governance



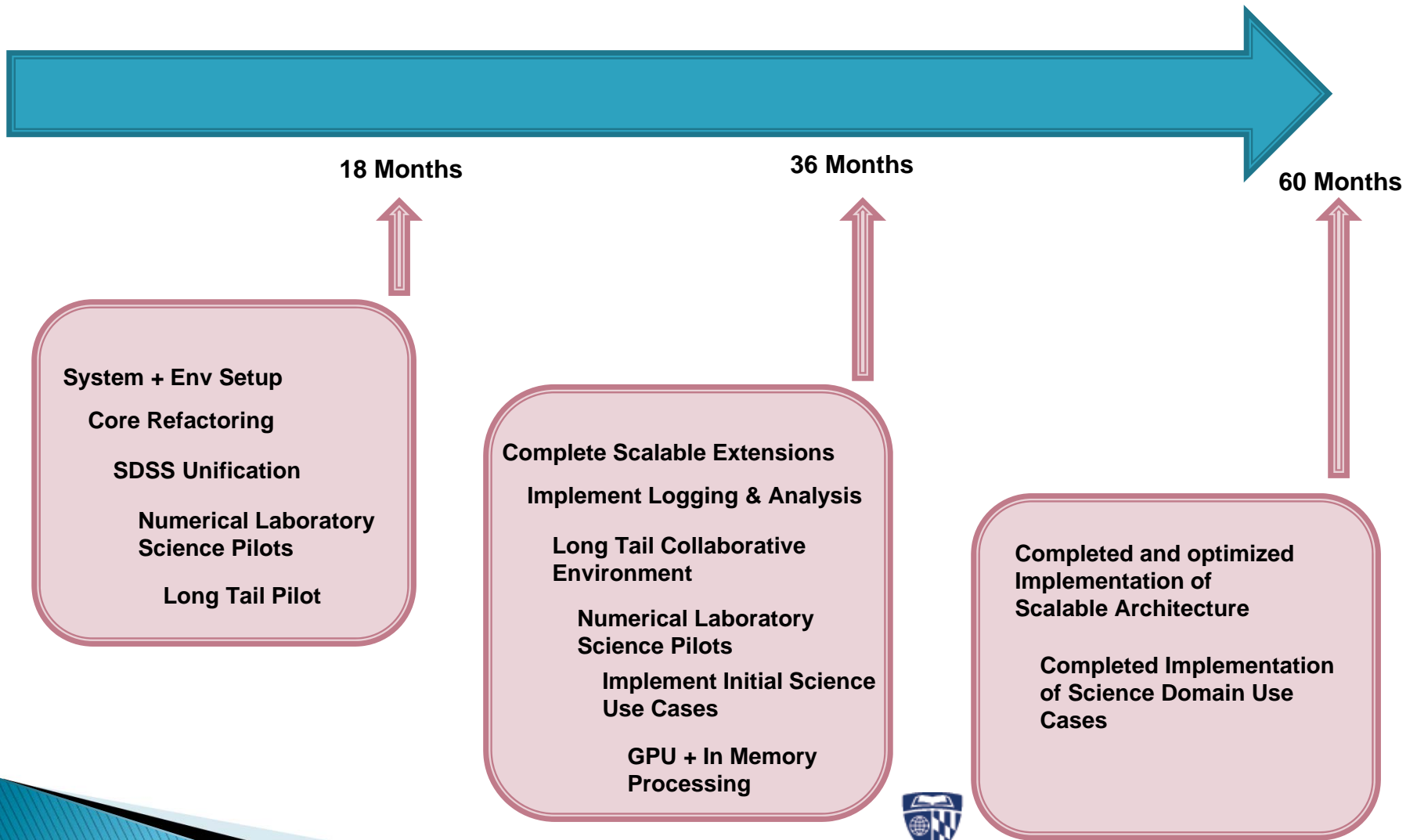
# Roadmap

- ▶ Structured around key Milestones:
  - 3 Month Planning (DONE)
  - 18 Month Review
  - 36 Month Review
  - 5 Year completion
- ▶ Defined at a high level in the PEP
- ▶ Is not expected to change at a high level, though details might with formal approved change requests

# High Level Plan



# High Level Roadmap



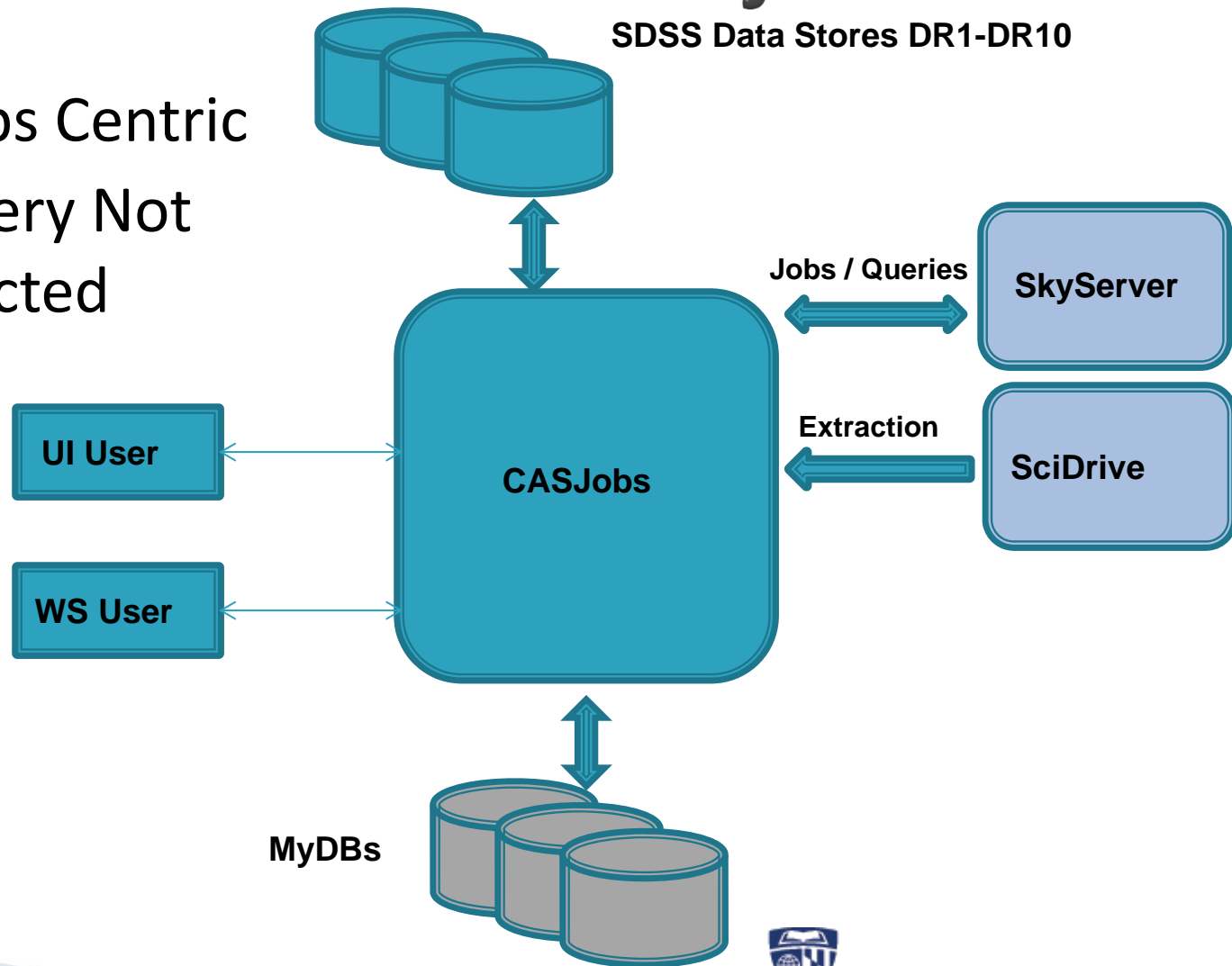


# Architectural Development

- ▶ ***Build off existing Building Blocks!***
- ▶ Development of SOA-like architecture to support modularization and interoperability
- ▶ Significant expansion of hardware infrastructure: servers, storage, GPUs
- ▶ Scalable architecture – GrayWulf cluster, database parallelization, App Layer VM pool
- ▶ Advanced processing through GPU and large in-memory platforms
- ▶ Cross-platform development and technologies: Windows, Linux; SQL-Server, PostGres; Java, .NET
- ▶ ***Scalability will be developed from the SkyQuery Architecture***

# Architecture – At Project Start

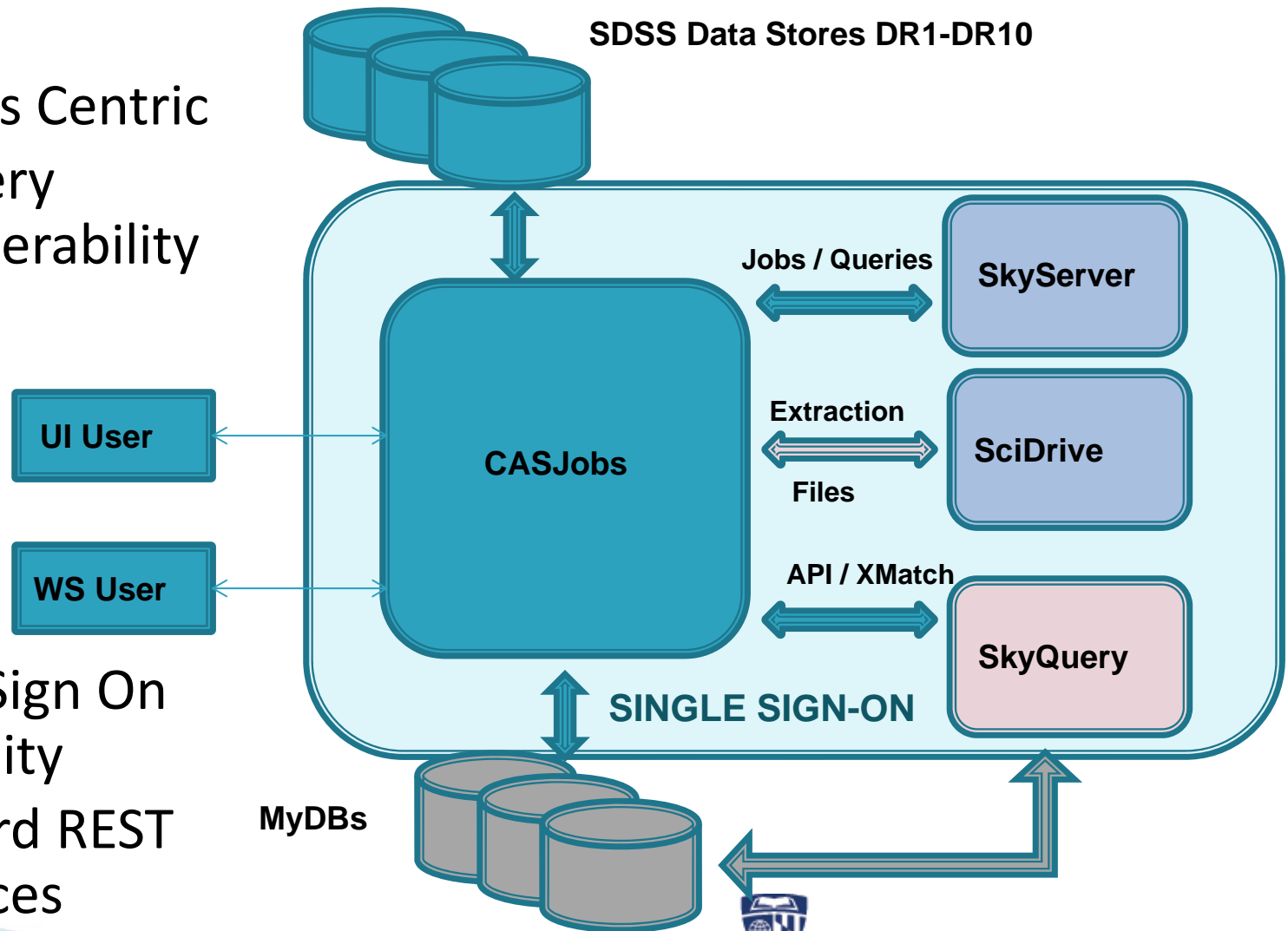
- ▶ CASJobs Centric
- ▶ SkyQuery Not Connected



# Architecture – Now

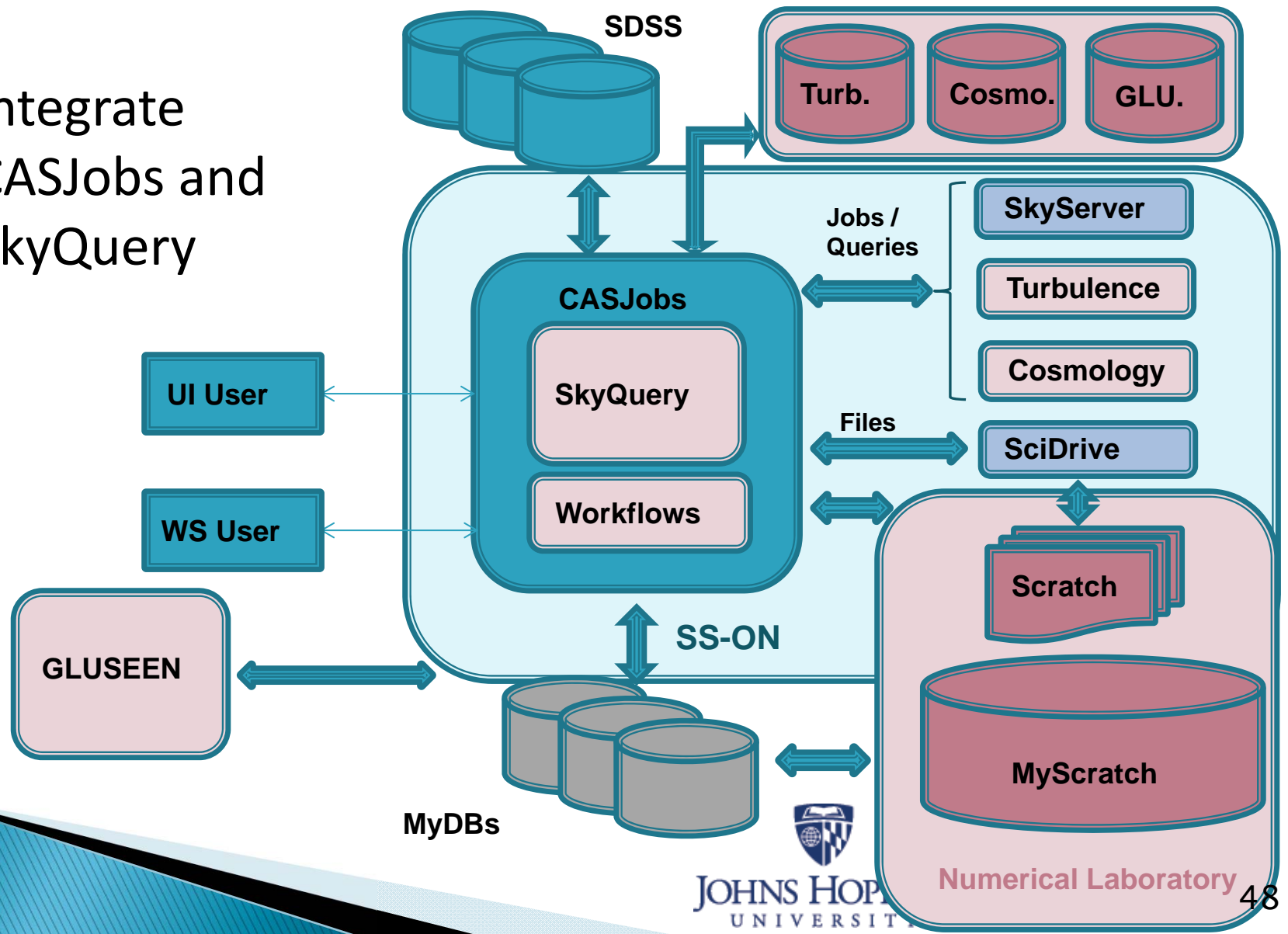
- ▶ CASJobs Centric
- ▶ SkyQuery Interoperability

- ▶ Single Sign On Capability
- ▶ Standard REST Interfaces



# Architecture – 18 Months

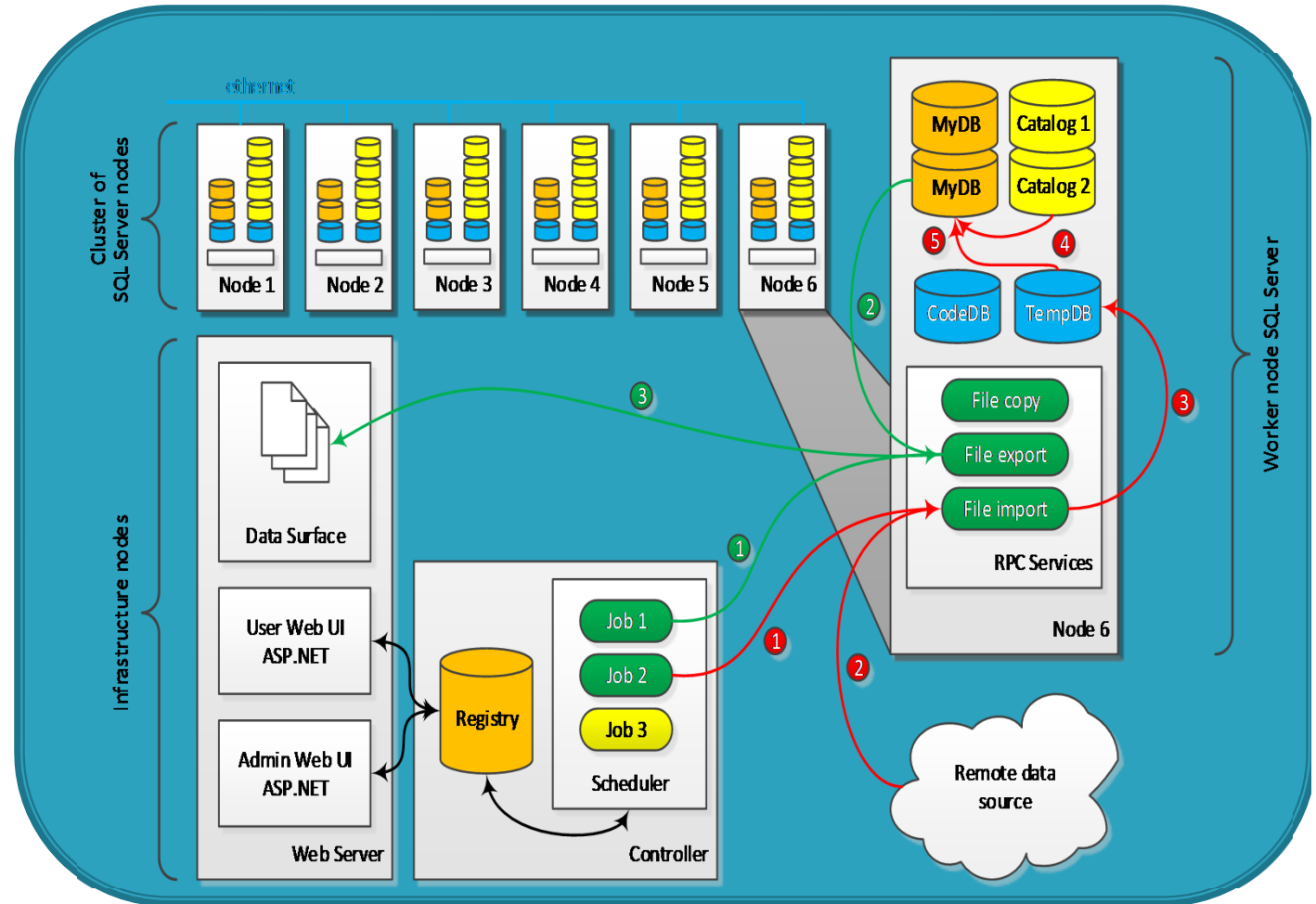
- ▶ Integrate CASJobs and SkyQuery



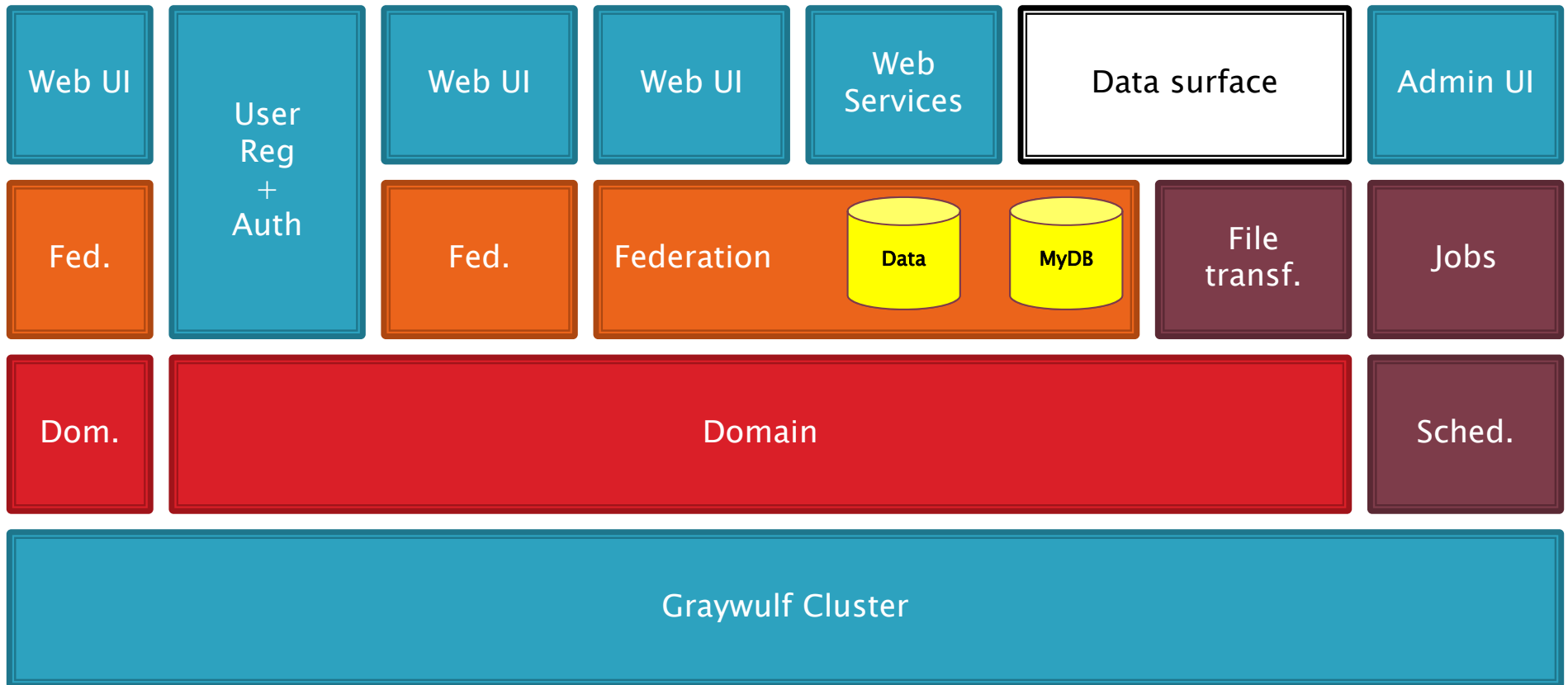


# Architecture – 3 Years

- ▶ Migrate to SkyQuery
- ▶ Extend Capabilities
- ▶ Support Application Tier



# SkyQuery Cluster Configuration

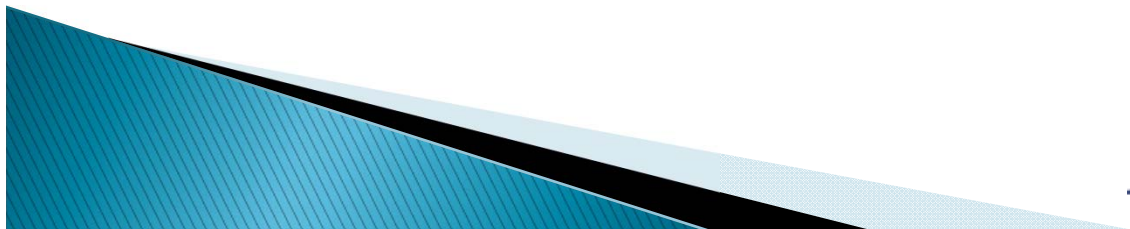


# Progress Overview (1)

- ▶ Planning + Team formation
- ▶ Initial Hardware purchase and upgrade
- ▶ SDSS (I, II and III) data unification almost complete
- ▶ Refactoring and modularization of the existing SkyServer infrastructure
- ▶ Standardization of Web Service API
- ▶ Central Authentication - Single Sign-on
- ▶ Upgrade of SciDrive “Dropbox-like” data storage tool

# Progress Overview (2)

- ▶ Initial Integration of SkyQuery and GrayWulf
- ▶ Initial development of more detailed Science Use Cases
- ▶ Design work to support Open Numerical Laboratory Pilot project using Turbulence simulation data
- ▶ Early development of GLUSEEN (Earth Science) use cases for Pilot implementation





# SDSS Data Migration Progress

- ▶ Unify SDSS-I/II/III/IV... websites
  - Done – new sdss.org launched on July 1, 2014
- ▶ Bring all SDSS-I and II data and services to JHU
  - All database (CAS) services at JHU as of Aug 1, 2014
  - Still need to move file access (DAS) services, although all the files have been downloaded (~ 80 TB)
- ▶ Consolidate Help Desks
  - Telecon held in mid-July to plan this

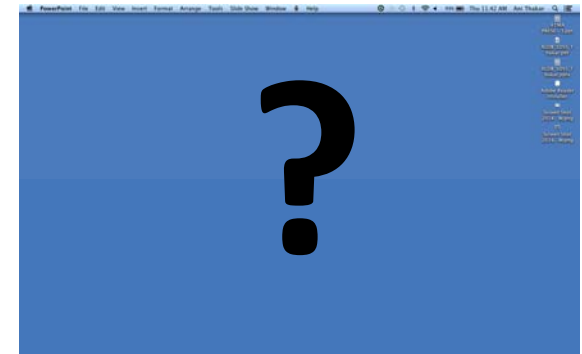
# Unified SDSS website



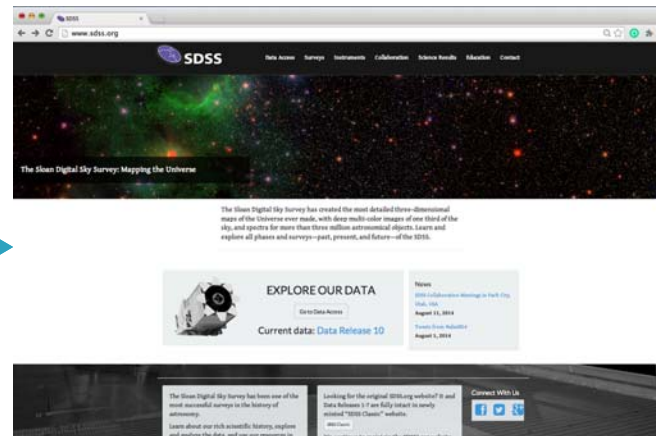
www.sdss.org



www.sdss3.org



www.sdss4.org

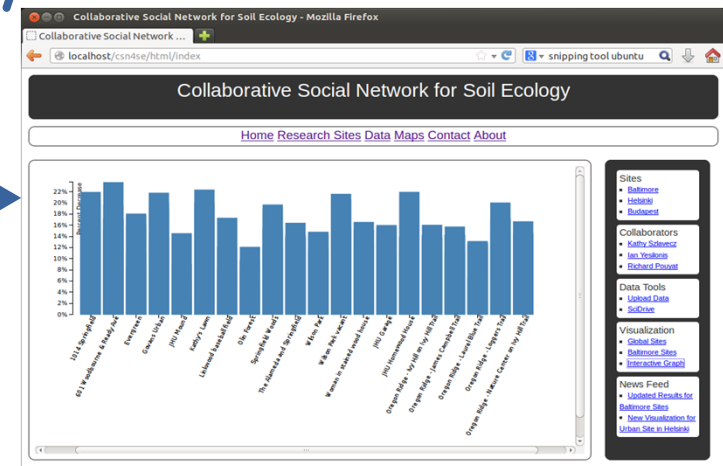
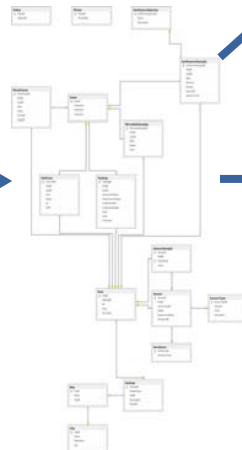
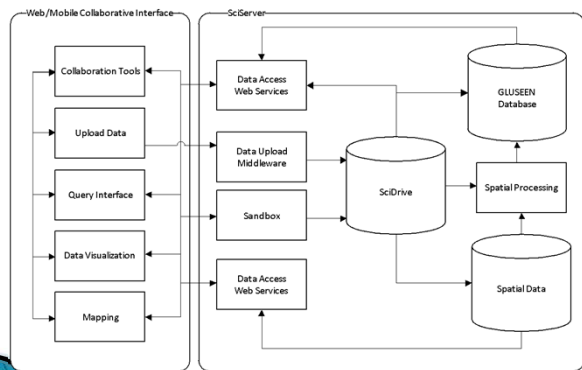


New [www.sdss.org](http://www.sdss.org) website

► Rebranding of SDSS web presence

# Progress on GLUSEEN

- ▶ High Level Use Cases
- ▶ Data Modeling, Schema Development
- ▶ Initial Data Load
- ▶ Early working UI proof of concept
- ▶ Design work for Pilot Use



# Process Improvement

- ▶ Things can always be done better
- ▶ Internal reviews
- ▶ Department wide
- ▶ Evaluating new Tools for better collaborative project and staff management
- ▶ Software engineering approach, tools and processes
- ▶ Maintain pragmatism and focus





# Branding and Website

- ▶ Name: SciServer
- ▶ Logo
- ▶ Website
  - In development
  - Wordpress CMS for easier editing by scientists



# Community Engagement

- ▶ “Science evangelism”
- ▶ Workshops at professional meetings
  - Science domains, info. sciences, education
- ▶ Online resources
  - Email helpdesk
  - Interactive tutorials
  - Stack Overflow-like system



# Training

## ► Multi-day summer schools

- Starting Year 3
- Modeled on Virtual Observatory's
- Online group follow-up



[home](#)  
[what is the nvo](#)  
[faq](#)  
[the nvo book](#)  
[behind the scenes](#)  
[documents](#)

The US National Virtual Observatory hosted its fourth Summer School on 3–11 September, 2008, at [The Lodge at Santa Fe](#), in Santa Fe, New Mexico. Forty-seven participants worked with experienced NVO scientists and software developers to learn how to do astrophysics with the Virtual Observatory. Participants also had the opportunity to work on self-motivated projects, building VO-enabled applications and doing VO-enabled research.

The US NVO Project greatly appreciates the sponsorship of NSF and NASA for the Summer School, and thanks the participants for their attention and interest in the Virtual Observatory.





# User Feedback

- ▶ Solicit feedback from data providers and end users
- ▶ User advisory group
  - Hold regular meetings & workshops
  - Group forming now, initial interviews complete
- ▶ External Advisory Board
  - Christine Borgman, Tony Hey, Dan Fay, Stu Feldman, Chris Mentzel, Dan Atkins
- ▶ Log (anonymous) usage stats from all websites and systems
  - Publish results

## computing in SCIENCE & ENGINEERING

### EXTREME DATA

- 8 Guest Editors' Introduction  
Manish Parashar and George K. Thiruvathukal  
Extreme Data
- 11 Big Data Applications Using Workflows for Data Parallel Computing  
Jianwu Wang, Daniel Crawl, Ilkay Altintas, and Weizhong Li  
In the Big Data era, workflow systems must embrace data parallel computing techniques for efficient data analysis and analytics. Here, an easy-to-use, scalable approach is presented to build and execute Big Data applications using actor-oriented modeling in data parallel computing. Two bioinformatics use cases for next-generation sequencing data analysis demonstrate the approach's feasibility.
- 22 Ten Years of SkyServer I: Tracking Web and SQL e-Science Usage  
M. Jordan Raddick, Ani R. Thakar, Alexander S. Szalay, and Rafael D.C. Santos  
SkyServer is the primary catalog data portal of the Sloan Digital Sky Survey that makes multiple terabytes of astronomy data available to the world. Here, the process is described of collecting and analyzing the complete record of more than 10 years of Web hits and SQL queries to SkyServer.
- 32 Ten Years of SkyServer II: How Astronomers and the Public Have Embraced e-Science  
M. Jordan Raddick, Ani R. Thakar, Alexander S. Szalay, and Rafael D.C. Santos



Cover illustration: Andrew Baker  
[www.debutart.com/illustration/andrew-baker](http://www.debutart.com/illustration/andrew-baker)

#### STATEMENT OF PURPOSE

*Computing in Science & Engineering* (CISE) aims to support and promote the emerging discipline of computational science and engineering and to foster the use of computers and computational techniques

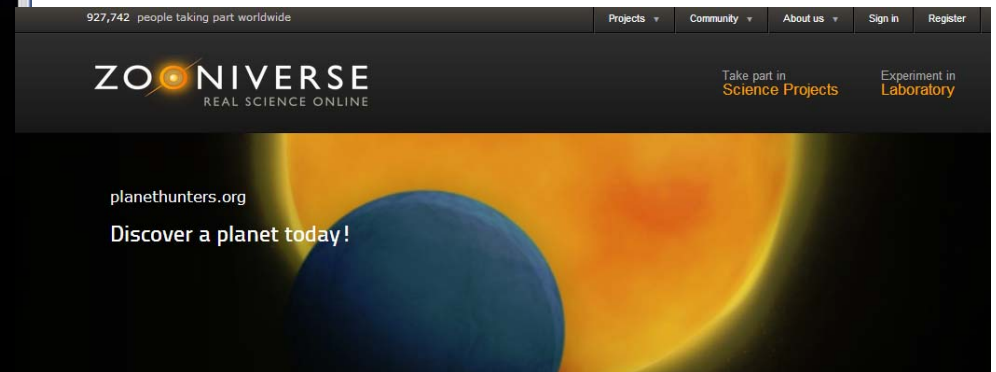
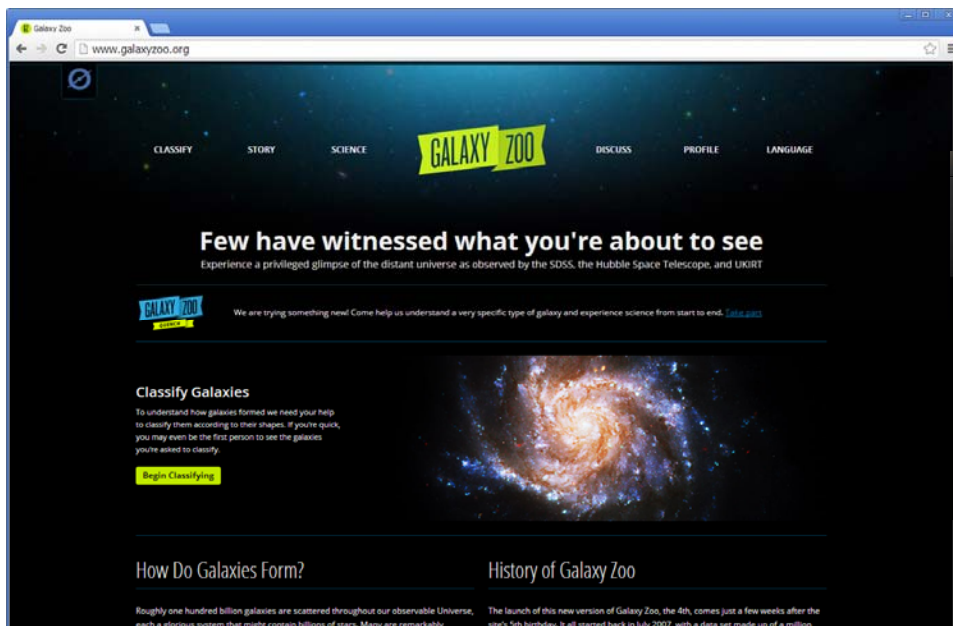


JOHNS HOPKINS  
UNIVERSITY



# Citizen Science

- ▶ Galaxy Zoo: citizen science with SDSS images
- ▶ Working with *Zooniverse* citizen science portal on other projects



# Student Notebook

- ▶ Design systems to include:
  - User logins and profiles (teachers and students)
  - **A student notebook** to store data, plots, and student responses online
  - **Teacher admin:** teachers can view their students' notebooks and query results
- ▶ Proof of concept:

The screenshot displays the SkyServer DR10 v2 web interface. The top navigation bar includes links for Home, Data, Schema, Education, Astronomy, SDSS, Contact Us, Download, Site Search, and Help. On the right, there are login and registration fields. The main content area shows the 'SkyServer NoteBook' interface in 'CasJobs Mode'. It features a table of astronomical data with columns for object ID, type, right ascension, declination, and various photometric bands. Each row of data includes links for 'Explore', 'Navigate', and 'Delete'.

**SLOAN DIGITAL SKY SURVEY III**  
**SkyServer DR10 v2**

Home Data Schema Education Astronomy SDSS Contact Us Download Site Search Help

Login:   
Password:

Current notebook: Default

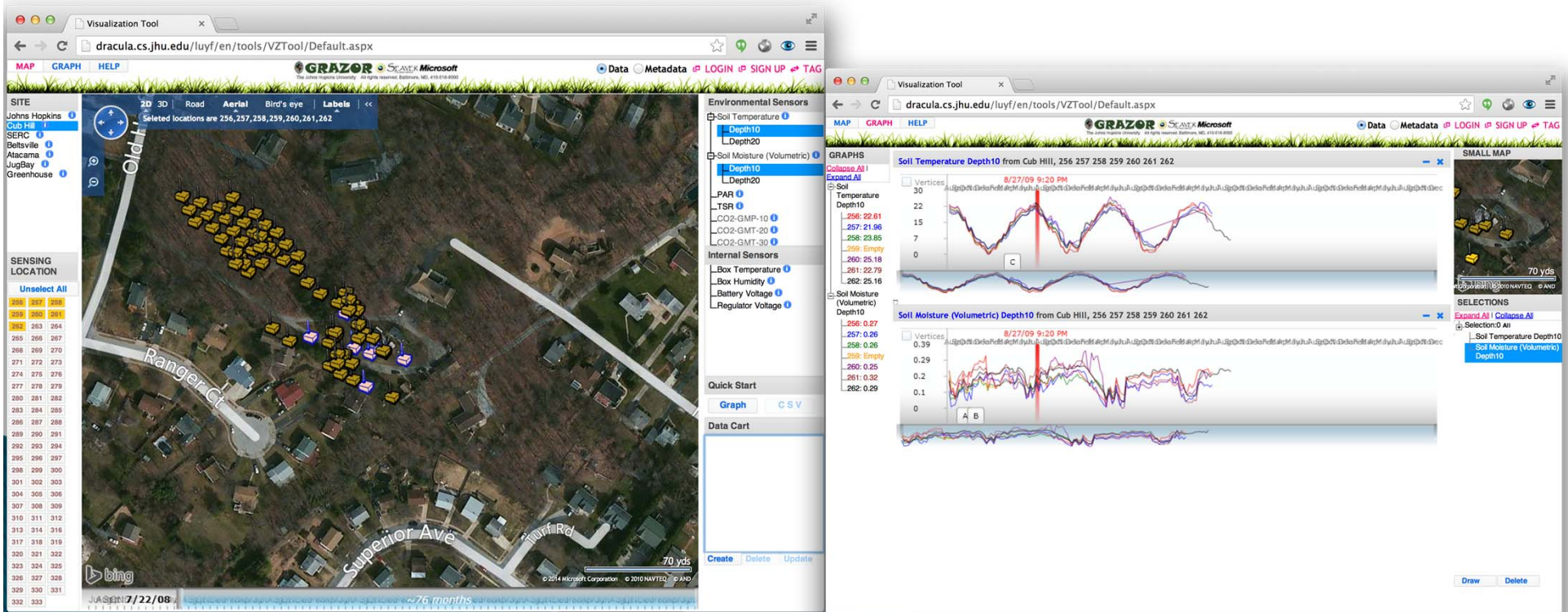
**SkyServer NoteBook**  
**CasJobs Mode**

objid	type	ra	dec	u	g	r	i	z	redshift			
1237661382772195356	STAR	145.271324	34.764381	15.46	13.61	13.16	14.71	13.82	-	<a href="#">Explore</a>	<a href="#">Navigate</a>	<a href="#">Delete</a>
1237661382772195474	GALAXY	145.267149	34.732906	17.28	15.47	14.63	14.15	13.79	-	<a href="#">Explore</a>	<a href="#">Navigate</a>	<a href="#">Delete</a>
1237661382772195480	GALAXY	145.265840	34.726143	25.06	21.49	23.87	23.49	21.99	-	<a href="#">Explore</a>	<a href="#">Navigate</a>	<a href="#">Delete</a>

▶ Full work starts Year 2

# Beyond SDSS SkyServer

- ▶ Expand to other sciences starting in Year 3
- ▶ Another example: “Grazor” viz tool for sensor data





# New Educational Activities

## ► Years 2-3: revise activities to use:

- Student notebooks
- Enhanced pedagogy
- Common Core science standards
- New data (SDSS, others)

## ► Years 3-5: expand to other domains

The screenshot shows the SkyServer DR 9 website. The top navigation bar includes 'SkyServer DR 9' and 'Tools'. Below this is a purple header with 'SDSS' and a navigation menu: 'Ground Control', 'Preflight Training', 'Launch', 'Expeditions', and 'Help'. The main content area is titled 'Redshift' and features a color gradient bar. It includes links for 'Required Preflight Training - SDSS Spectrum Graphs' and 'Recommended Preflight Training - Redshift'. A sidebar on the left has buttons for 'Launch into SDSS', 'Launch Solar System', 'Launch Milky Way', and 'Launch Cosmos'. The text explains that with basic understanding of redshift and the SDSS spectrum graph tool, users can explore how redshift is measured. It also introduces the Science Archive Server (SAS) as a place to access spectra. Two buttons, 'I Have a Starting Place' and 'Use Constellations Notebook', are provided. A paragraph explains how to navigate the SAS Spectra List page, noting that each row represents a different object and that columns for Survey, Plate ID, and MJD are identical. A footer bar contains 'DR9 Science Archive Server (SAS)' and a 'Tell Me More' button.

SkyServer DR 9 Tools

SDSS Ground Control Preflight Training Launch Expeditions Help

### Redshift

Required [Preflight Training - SDSS Spectrum Graphs](#)  
Recommended [Preflight Training - Redshift](#)

With some basic understanding about redshift and the tool of the SDSS spectrum graph in hand, you are prepared to explore how redshift is measured and how it can be used. We will need lots of spectra.

### Accessing Data Through the Science Archive Server (SAS)

If you want to look at a lot of spectra at one time, the easiest place to access them is through the Science Archive Server (SAS). SAS is the latest image and spectrum service for the SDSS. But before you can use SAS, you need a starting place. Choose one of the paths below. If you already have a [Special Place in the Database](#), great. Start there. If you don't, choose one from the SDSS Constellations Notebook.

[I Have a Starting Place](#) [Use Constellations Notebook](#)

When you arrive at a SAS Spectra List page, bookmark the location. Next, notice that each row represents a different object that was captured by the spectrograph. The Survey, Plate ID, and MJD columns are identical. This set of data was gathered under the same observing goals ([Survey](#)) using the same [spectroscopic plate](#) (Plate) on the same day ([MJD](#)). It isn't until you get to the fourth column (Fiber #) that the information becomes unique. Scroll down and notice that there are either 640 or 1000 objects in the list depending upon the survey. You can reorder any column by clicking the up-down arrows on the column heading.

DR9 Science Archive Server (SAS)

Home Spectra Images Documentation Tell Me More

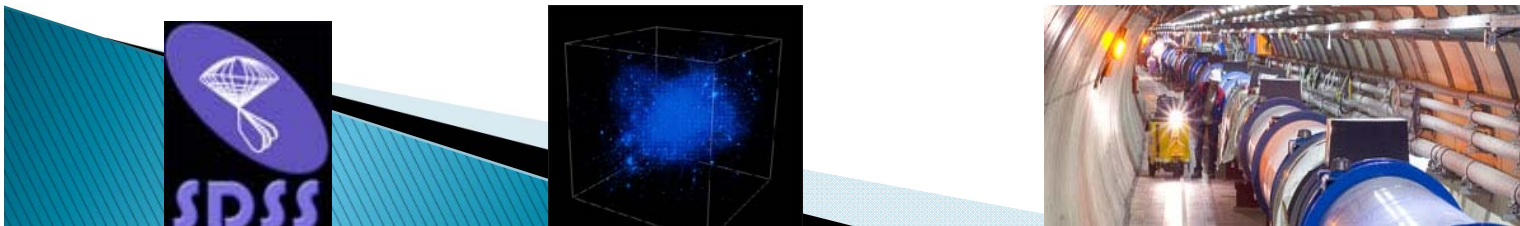
# DIBBs Partner Collaboration

- ▶ Purdue University – “Integrating Geospatial Capabilities into HUBzero” (GABBS)
  - Synergy: Geospatial Data Integration in Long Tail
- ▶ Carnegie-Mellon University – “The Data Exacell” (Astronomy overlap)
  - Synergy: Simulations Into Numerical Laboratory
- ▶ University of Illinois – “Brown Dog”
  - Synergy: Migration Tools



# Trends

- ▶ Broad sociological changes
  - Convergence of Physical and Life Sciences
  - Data collection in ever larger collaborations
  - Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,...
  - Analysis decoupled, off archived data by smaller groups
  - Emergence of the citizen/internet scientist (GalaxyZoo...)
- ▶ Need to start training the next generations
  - $\Pi$ -shaped vs I- and T-shaped people
  - Early involvement in “Computational thinking”



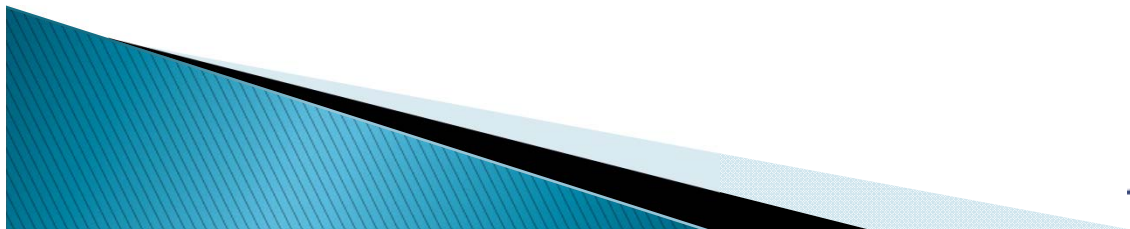
# Summary

- ▶ Science is increasingly driven by data (big and small)
- ▶ Surveys analyzed by individuals
- ▶ From hypothesis-driven to data-driven science
- ▶ New instruments: “microscopes” & “telescopes” for data
- ▶ A major challenge on the “long tail”
- ▶ A new, Fourth Paradigm of Science is emerging...
- ▶ SDSS has been at the cusp of this transition
- ▶ Now the SciServer is continuing the Jim Gray legacy



*“If I had asked people what they wanted, they would have said faster horses...”*

—Henry Ford



# Contacts

- ▶ Alex Szalay ([szalay@jhu.edu](mailto:szalay@jhu.edu))
- ▶ Ani Thakar ([thakar@pha.jhu.edu](mailto:thakar@pha.jhu.edu))
- ▶ Jordan Raddick ([raddick@jhu.edu](mailto:raddick@jhu.edu))
- ▶ Mike Rippin ([mike.rippin@jhu.edu](mailto:mike.rippin@jhu.edu))