Science Platforms

Ani Thakar
JHU Institute for Data Intensive Engineering and Science (IDIES)
LSSTC-DSFP Workshop, March 2019

Science Platforms

- Emerging consensus on science platforms
- Definition as per 2018 STScI Workshop:
  - The term "Science Platform" is growing in popularity and is generally used to refer to an integrated system of web applications, programmatic services (Application Programming Interfaces, or APIs), and an interactive analysis environment that are connected to a collection of astronomical data services.
- Server-Side Analysis is key
  - Upload (and execute) analyses close to large datasets

Science Platform Wishlist

- Authentication and single sign-on
- Interactive server-side analysis environment
  - Jupyter based, Python/R/Matlab/Julia/…
  - Can you create new images?
- Query services
  - Synchronous (browser-based) query
  - Asynchronous (batch mode) query
  - Ability to deal with large query outputs

Science Platform Wishlist

- **Spectrum services**
  - Query spectra from large datasets
  - Interactive plot/visualize spectra
  - Analyze spectra
- **Image cutout and mosaicking service**
  - FITS cutout, preferably
  - Can it handle arbitrarily large mosaics?
  - Finding chart feature
  - Interactive browsing of image data

Science Platform Wishlist

- **Cross match service**
  - Unlimited cross-matching between data sets
  - Upload your own data to cross-match
- **Collaboration features: groups and access control**
  - Ability to create and administer groups
  - Share resources with collaborators, students
  - Give permissions as necessary to group members
- **File management**
  - Upload and download file-based data, extract metadata

Science Platform Wishlist

- Job management services
  - Asynchronous (batch) job management
  - Job history and session management
  - Fair use of available resources (like queues)
- Help desk
  - Responsive on time scale of a day or two at least
- Backup services – is my data backed up?
- Support VO protocols
  - TAP, SIAP, SSAP, …

Questions

What would you add to Science Platform requirements?

- What's missing?
- Your favorite tool(s) or features

What are the current science platforms out there?

- Compare them
- Evaluate merits
- Rate them

Science Platform: SciServer

- A Science Platform across science domains
- NSF DIBBs (Data Infrastructure Building Blocks)
  - Leverage SDSS building blocks (SkyServer and CasJobs)
  - Extend and generalize them for versatile science platform
  - Enable collaborative data-driven science
- The official science platform for SDSS catalog data

SciServer Team

Gerard Lemson (SciServer Compute, Science Lead)

Mike Rippin (SciServer Project Manager)

JaiWon Kim (Web dev, CourseWare)

Bonnie Souter (Website, Github)

Jordan Raddick (SkyServer, Websites, EPO)

Manu Taghizadeh-Popp (SkyServer, SciServer, SciQuery)

Dmitry Medvedev (CasJobs, SciServer)

Camy Chhetri (SciServer Dashboard, CourseWare)

Joseph Booker (SciServer, Compute)

Sue Werner (SQL Server) Rutgers CS, Class of 2000

Mehwish Zuberi (DBA)

Victor Paul (Senior DBA)

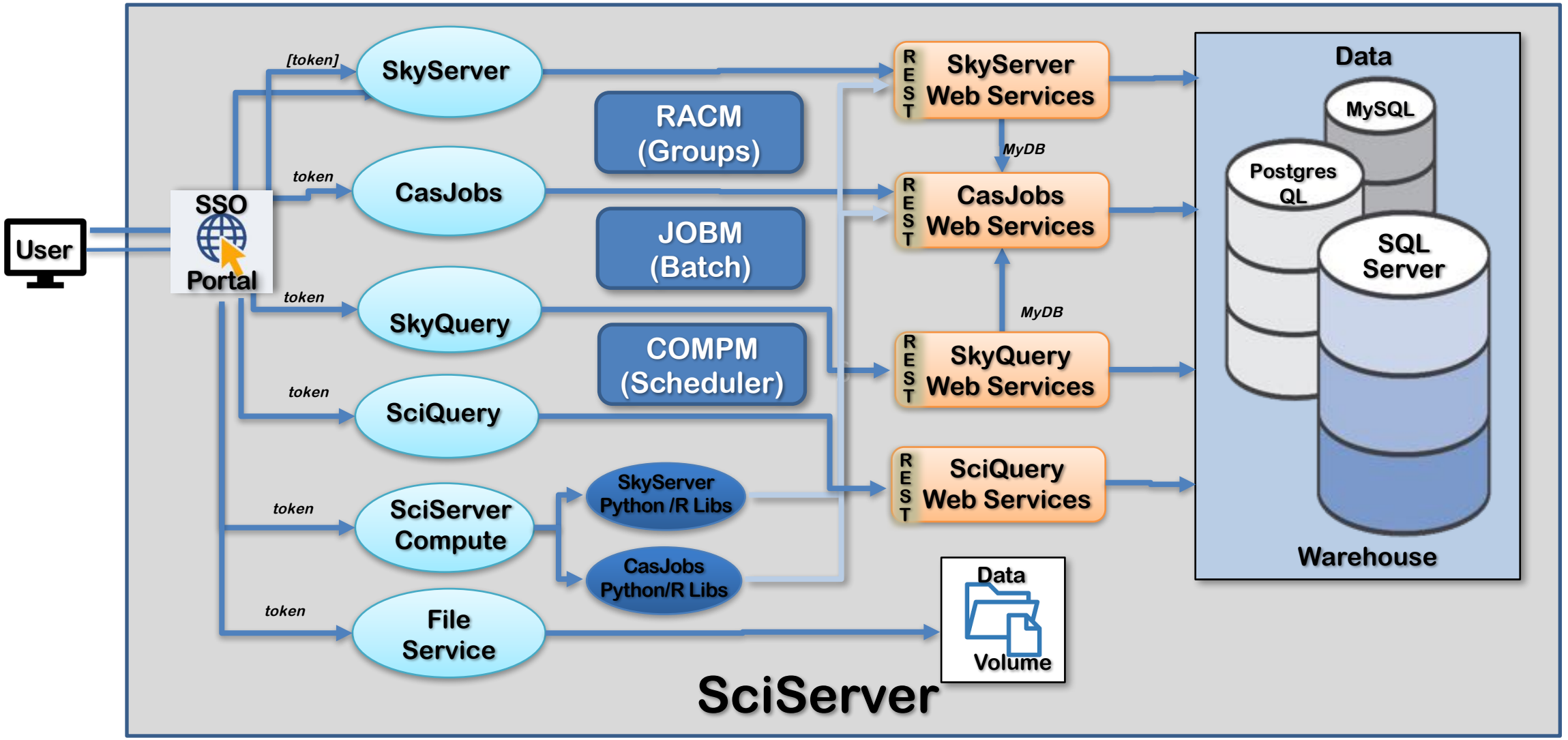
Lance Joseph (IT Nerd)

Jan vandenBerg (IT Manager)

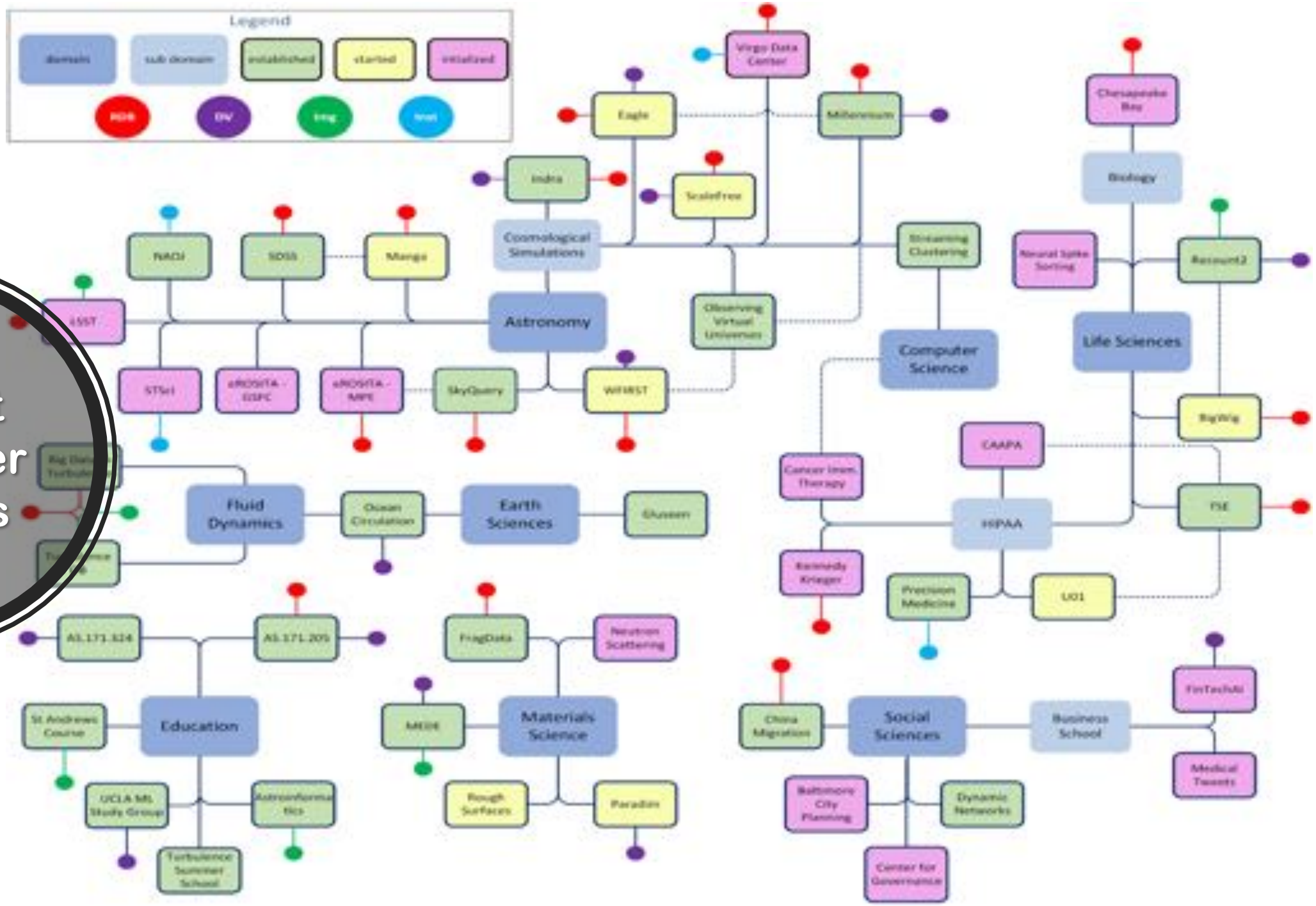
SciServer Features

- Single sign-on
- Dashboard of services
  - SkyServer, CasJobs, Compute, Files, Groups, SkyQuery
  - Coming soon: SciQuery and CourseWare
    - Although classroom features already pretty good
- Collaborative Science Framework
- Compute is the most powerful component
  - Jupyter notebook environment combined with asynchronous job management and resource sharing
    - Python/R/Matlab/Julia
  - Docker/VM containerized architecture
  - Custom docker images for specialized libraries and applications
- Web Services architecture with REST APIs to all services

SciServer Architecture



Current SciServer Projects



SciServer Tutorial

http://sciserver.org/

Exercises

1. Create a new python notebook that includes
   a) The CROSS APPLY query that you created yesterday to find neighbors of the selected objects
   b) Modify the query to obtain the distance between each object and its neighbor (hint: use a user-defined function for this)
   c) Run the query on CasJobs in batch mode
      a) Read the resulting MyDB table into the notebook
   d) Get JPEG cutouts for all the objects from SkyServer, with each cutout big enough to include the neighbors you found
   e) After you are done, save the notebook in your Storage folder
2. Compare SciServer with other science platforms
   a) Evaluate comparative merits, features, what's missing