ADA MASTER PROGRAM:

BIG DATA IN ASTRONOMY TUTORIAL 3

Dragana Ilić
University of Belgrade – Faculty of Mathematics

https://github.com/ilicdragana/ADA.BigDataAstro

МАТФ
University of Belgrade
Faculty of Mathematics

150 година
МАТФ
Универзитет у Београду
Математички факултет

VIRTUAL OBSERVATORY CONCEPT

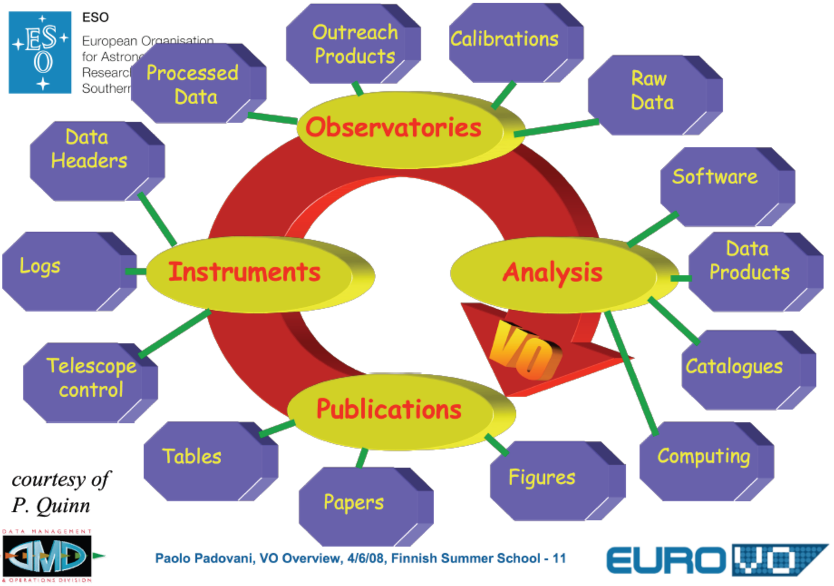
The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole.

The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible.

Important aims: define standards, develop software

https://www.ivoa.net/documents/

https://www.euro-vo.org/software/



courtesy of P. Quinn

Paolo Padovani, VO Overview, 4/6/08, Finnish Summer School - 11

VO TABELS

Virtual Observatory (**VO**) **tables** - based on *Extensible Markup Language (XML)*

metadata – parametri, opisi, razne informacije…

VOTable *: Standard format for representing and exchanging tabular data within the VO. Most data archives now offer export of data in VOTable format, and a large number of tools read VOTables. Any FITS table can also be expressed as a VOTable, but not always the other way round, because VOTable can hold more structured metadata, rather than just keyword-value pairs.*

http://ivoa.net/astronomers/vo_glossary.html

DATA FORMATS USED IN ASTRONOMY BIG DATA

FITS - Flexible Image Transport System

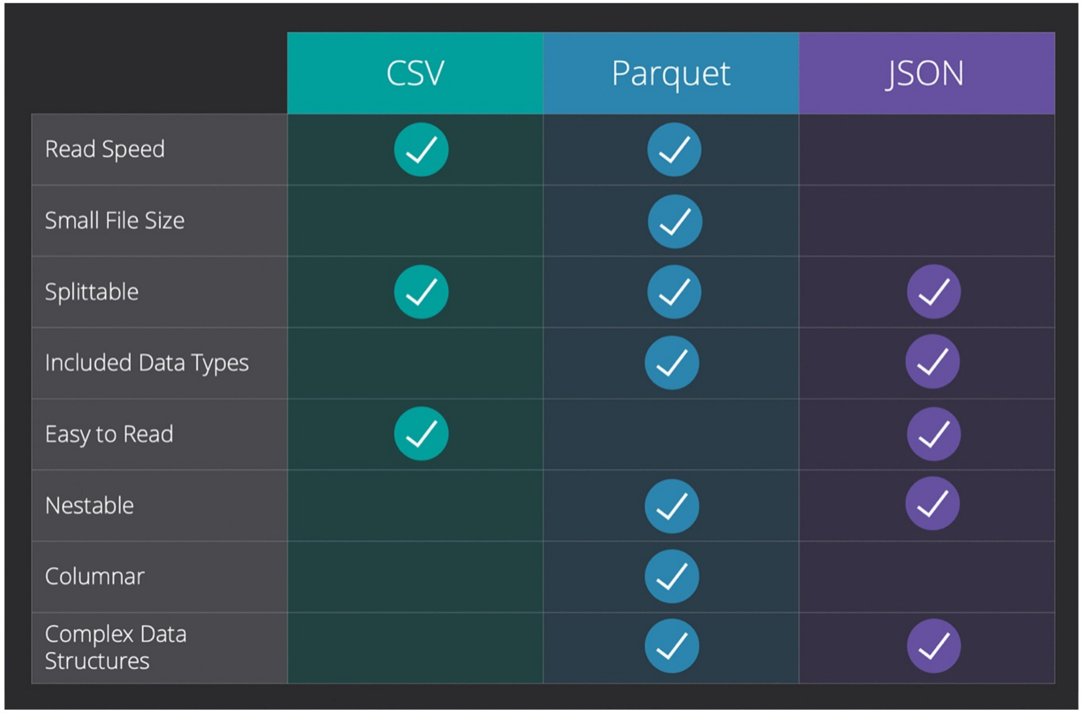
- one or more Header + Data Units (HDUs)
- formatted as multi-dimensional arrays (for example a 2D image), or tables
- most commonly used digital file format in astronomy.

VO Tables

cvs - Comma Separated Values – plain text

json - JavaScript Object Notation

Parquet - binary column oriented schema-aware file format



	CSV	Parquet	JSON
Read Speed	✓	✓	
Small File Size		✓	
Splittable	✓	✓	✓
Included Data Types		✓	✓
Easy to Read	✓		✓
Nestable		✓	✓
Columnar		✓	
Complex Data Structures		✓	✓

https://weber-stephen.medium.com/csv-vs-parquet-vs-json-for-data-science-cf3733175176

SciServer

Collaborative data-driven science

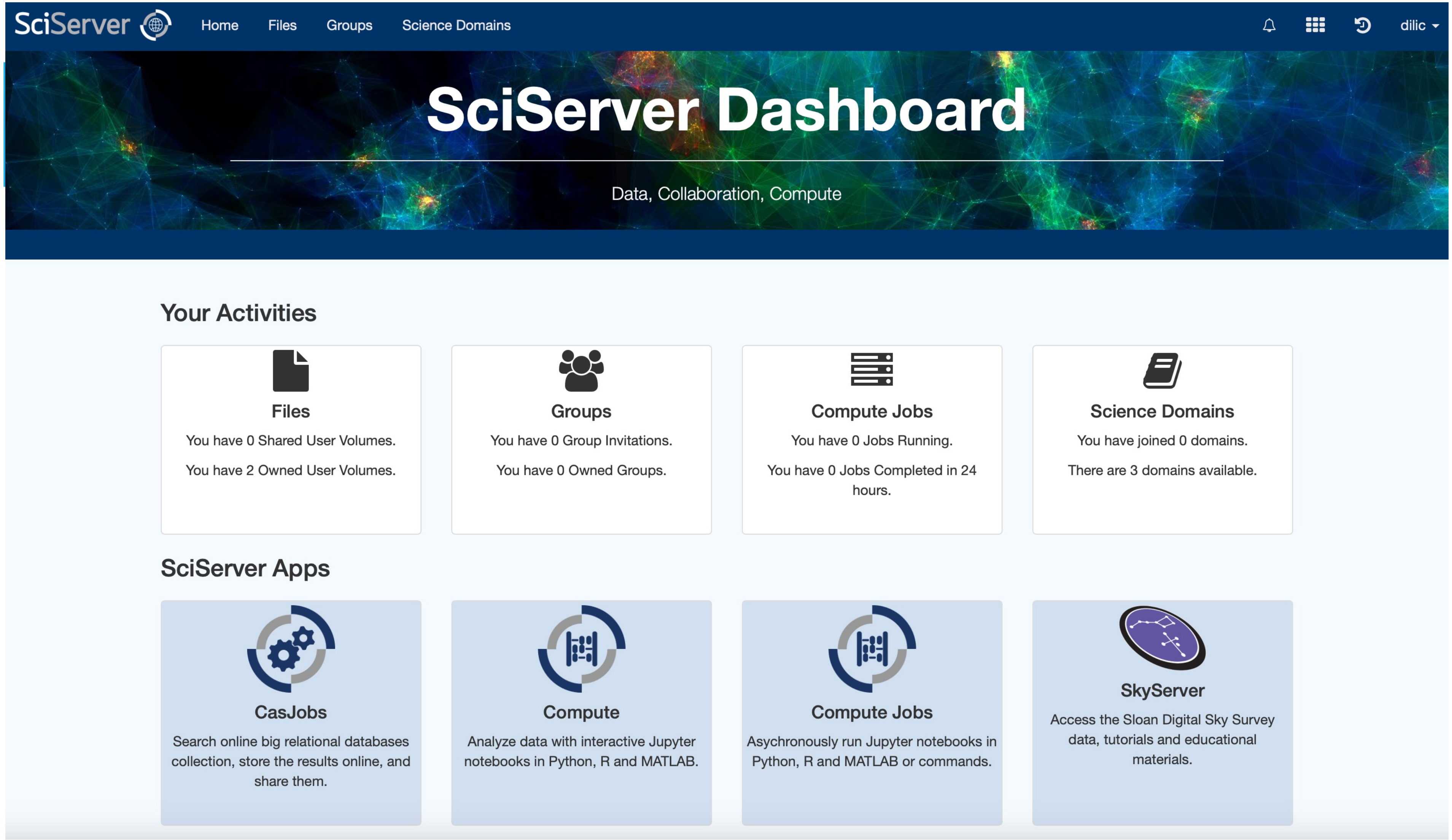
The Science Platform

A collaborative environment for server-side analysis with extremely large datasets

SciServer *Betelgeuse* v2.1.0

https://www.sciserver.org

- revolutionary new approach to doing science by bringing the analysis to the data
- consists of data hosting services coupled with integrated Tools that work together to create a full-featured system.
- operated by the Institute for Data Intensive Engineering and Science at Johns Hopkins University
- Hosts SkyServer - the primary public interface to catalog data from the Sloan Digital Sky Survey (SDSS), as well as CasJob to query the data



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.

Science Domains

You have joined 0 domains.

There are 3 domains available.

SciServer Apps

CasJobs

Search online big relational databases collection, store the results online, and share them.

Compute

Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.

Compute Jobs

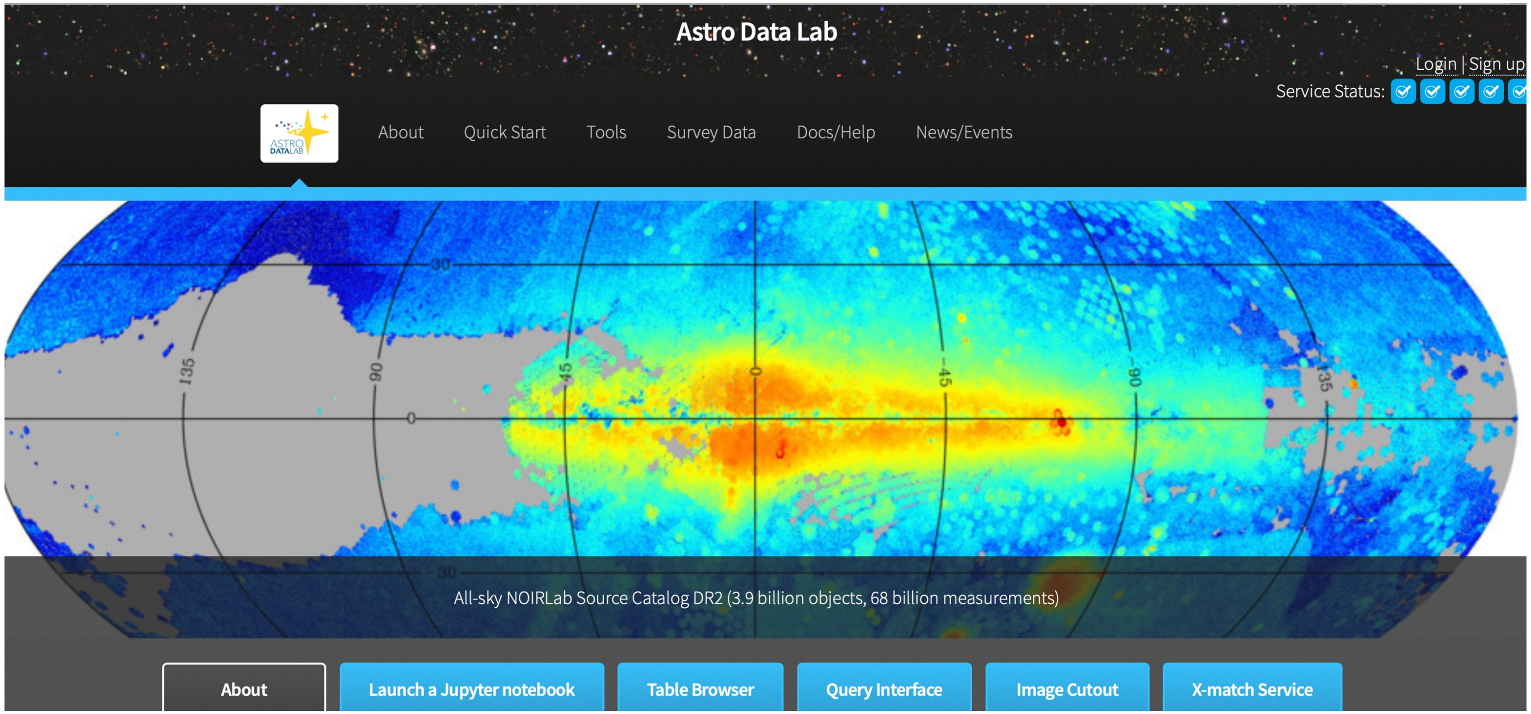
Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.

SkyServer

Access the Sloan Digital Sky Survey data, tutorials and educational materials.

NOIR LAB – ASTRO DATA LAB

Open an account at https://datalab.noirlab.edu

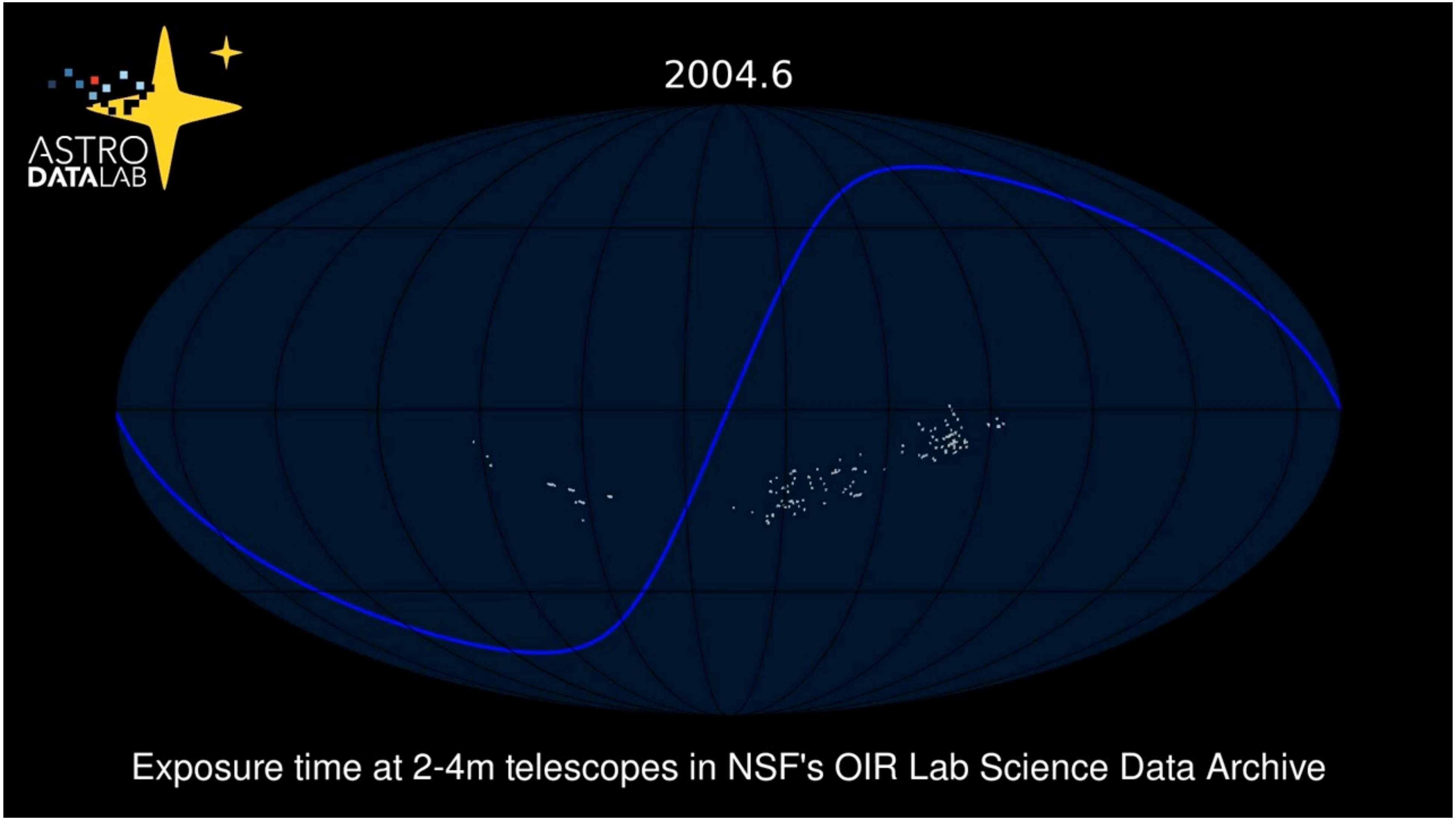


ASTRO DATA LAB AIMS

The overall goal of Astro Data Lab is to enable efficient exploration and analysis of the large datasets now being generated by instruments on NOIRLab and other wide-field telescopes.

- Connect users to high-value catalogs from NOIRLab and external sources (e.g. SDSS, GAIA) and NOIRLab-based images linked to catalog objects
- Enable users to discover the data that they need for their science
- Allow users to develop intuition through interaction with selected catalog and image sets
- Allow users to automate their analysis to aid discovery in large datasets

Astro Data Lab provides services to enable as much work as possible close to the data, while allowing transfer of data and results to local hardware anytime during the process.



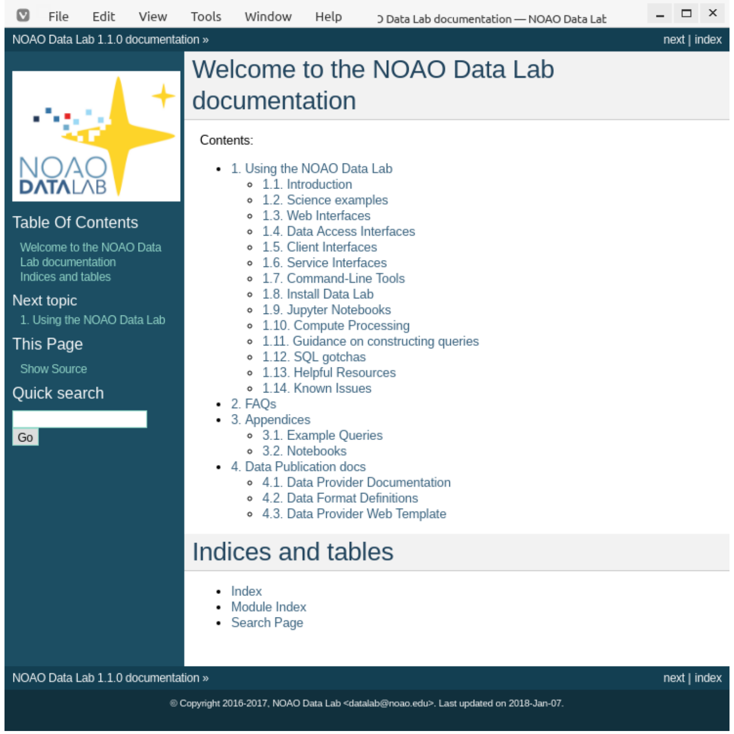
2004.6

Exposure time at 2-4m telescopes in NSF's OIR Lab Science Data Archive

ASTRO DATA LAB USERS

User = someone with a DL account

- Get 1 TB of storage on vospace
- Get 250 GB of MyDB storage
- Can upload datasets for joint analysis
- Can edit/create/delete notebooks
- Can upload own Python source code
- Can share data with others


File Edit View Tools Window Help
Data Lab documentation — NOAO Data Lab
NOAO Data Lab 1.1.0 documentation »
next | index
NOAO DATALAB

Welcome to the NOAO Data Lab documentation

Contents:

- 1. Using the NOAO Data Lab
  - 1.1. Introduction
  - 1.2. Science examples
  - 1.3. Web Interfaces
  - 1.4. Data Access Interfaces
  - 1.5. Client Interfaces
  - 1.6. Service Interfaces
  - 1.7. Command-Line Tools
  - 1.8. Install Data Lab
  - 1.9. Jupyter Notebooks
  - 1.10. Compute Processing
  - 1.11. Guidance on constructing queries
  - 1.12. SQL gotchas
  - 1.13. Helpful Resources
  - 1.14. Known Issues
- 2. FAQs
- 3. Appendices
  - 3.1. Example Queries
  - 3.2. Notebooks
- 4. Data Publication docs
  - 4.1. Data Provider Documentation
  - 4.2. Data Format Definitions
  - 4.3. Data Provider Web Template

Indices and tables

- Index
- Module Index
- Search Page

Table Of Contents

Welcome to the NOAO Data Lab documentation
Indices and tables

Next topic

1. Using the NOAO Data Lab

This Page

Show Source

Quick search

Go

NOAO Data Lab 1.1.0 documentation » next | index

© Copyright 2016-2017, NOAO Data Lab <datalab@noao.edu>. Last updated on 2018-Jan-07.

Overviews, glossary, science cases, SQL examples, tips & tricks

Curated default notebooks, and many science examples

User Forum & Manual - https://datalab.noirlab.edu/docs/manual/index.html

How to copy latest Data Lab notebooks to a new (writable) directory
https://datalab.noao.edu/docs/manual/UsingTheNOAODataLab/JupyterNotebooks/JupyterNotebooks.html#get-the-latest-set-of-default-notebooks

EXAMPLE NOTEBOOKS

Example 1: Getting Started

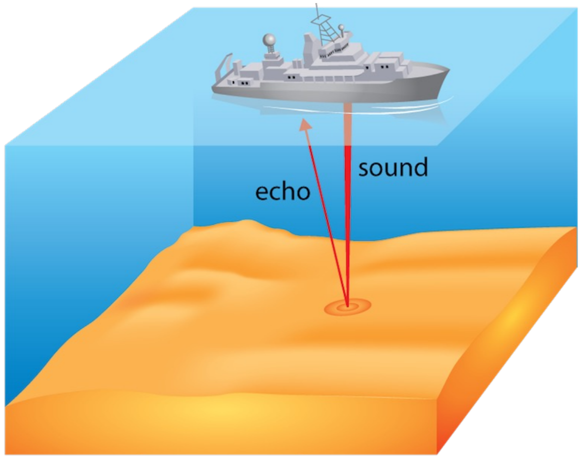
notebooks-latest/01_GettingStartedWithDataLab/02_GettingStartedWithDataLab.ipynb

Example 2: Echo-mapping of AGN

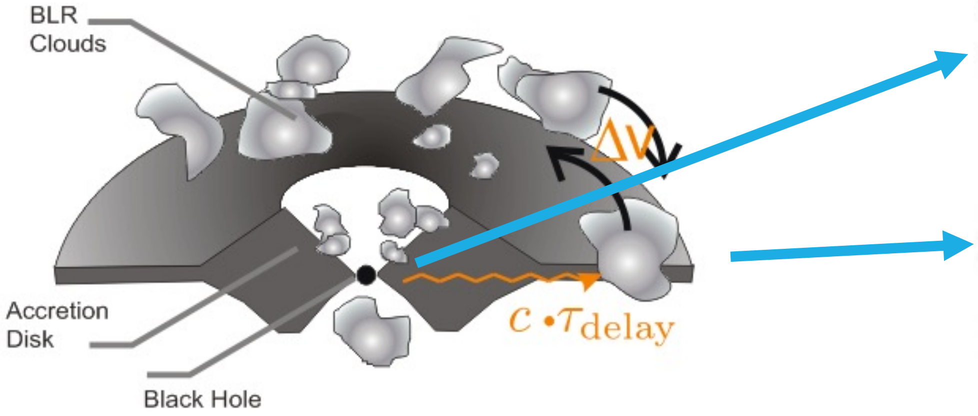
notebooks-latest/05_Contrib/TimeDomain/PhotoReverberationMappingAGN

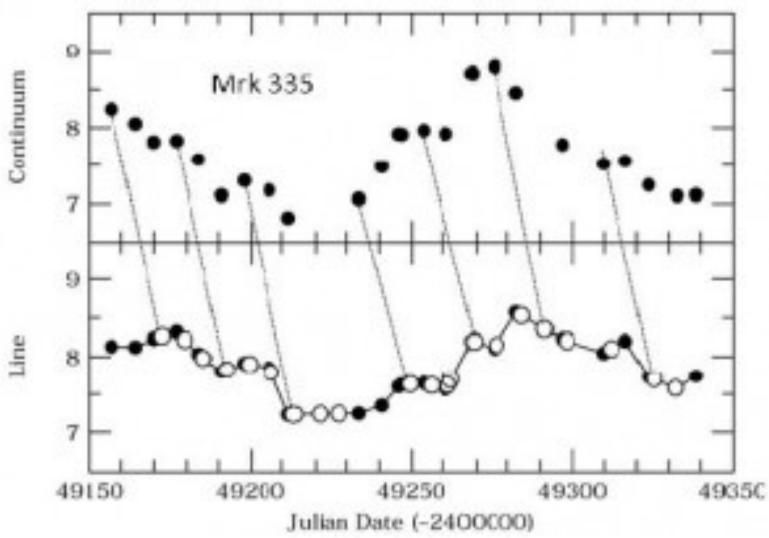
(see next slide for science case summary)

REVERBERATION MAPPING



- reverberation (echo) mapping in active galactic nuclei (called AGN or quasars) → way to study the "invisible" (unresolvable) galactic centers using time-domain astronomy
- we measure time-delay (using cross-correlation analysis) between the signal in the continuum and emission lines





Astrobites: https://astrobites.org/2012/03/14/measuring-the-black-hole-mass-in-markarian-6-using-reverberation-mapping/

Peterson 2001

SOFTWARE AND TIPS

- Course github page: https://github.com/ilicdragana/ADA.BigDataAstro
- Open github account for reports

- Python and Jupyter Notebooks

- Tip: create a new conda environment for this course:
  - conda create -n bda python=3.10
  - conda activate bda
  - conda install numpy matplotlib pandas seaborn scipy notebook
- For Jupyter notebook, type in terminal: jupyter-notebook

- Python Crash Course – by I. Jankov (PhD Student of Astronomy&Astrophysics) https://github.com/cefeida42/mass-agn/tree/b2979d0728d86ec86bab524f3c3362ffed03c569/Tutorial%200%20-%20Python%20Crash%20Course
  - Python syntax refresher
  - NumPy basics
  - Generating plots with matplotlib
  - Manipulating data with pandas

ASTRO HELP

https://pandas.pydata.org/docs/index.html

https://docs.scipy.org/doc/scipy/reference/main_namespace.html

https://colorbrewer2.org/

https://www.astropy.org

https://ui.adsabs.harvard.edu

https://datalab.noirlab.edu

https://www.astroml.org

astroML INTERACTIVE BOOK

- astroML is a Python module for machine learning and data mining that accompanies the book **"Statistics, Data Mining, and Machine Learning in Astronomy"** by Željko Ivezić, Andrew Connolly, Jacob Vanderplas, and Alex Gray.
- astroML is built on numpy, scipy, scikit-learn, matplotlib, and astropy → growing library of statistical and machine learning routines for analyzing astronomical data.
- notebooks that describe the statistical and machine learning methods used in astroML together with code that runs these methods on existing astronomical data sets
  - structure follows the chapters of the book
  - each notebook can viewed through the browser (with navigation links at the side of the page) or be downloaded to your own computer, or be executed directly using Binder or Google Colab

QUESTIONS

Email: dragana.ilic@matf.bg.ac.rs

Course github page: https://github.com/ilicdragana/ADA.BigDataAstro