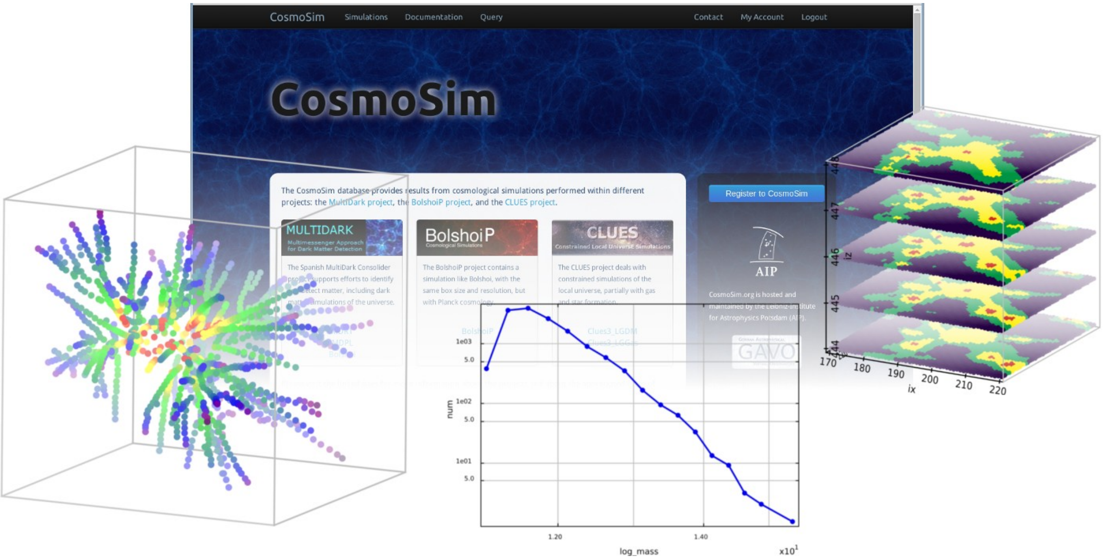
CosmoSim Database



The CosmoSim database provides results from cosmological simulations performed within different projects: the MultiDark project, the BolshoiP project, and the CLUES project.

The Spanish MultiDark Consolider project supports efforts to identify detect matter, including dark matter simulations of the universe.

The BolshoiP project contains a simulation like Bolshoi, with the same box size and resolution, but with Planck cosmology.

The CLUES project deals with constrained simulations of the local universe, partially with gas and star formation.

CosmoSim.org is hosted and maintained by the Leibniz-Institute for Astrophysics Potsdam (AIP).

Kristin Riebe
E-Science group @AIP

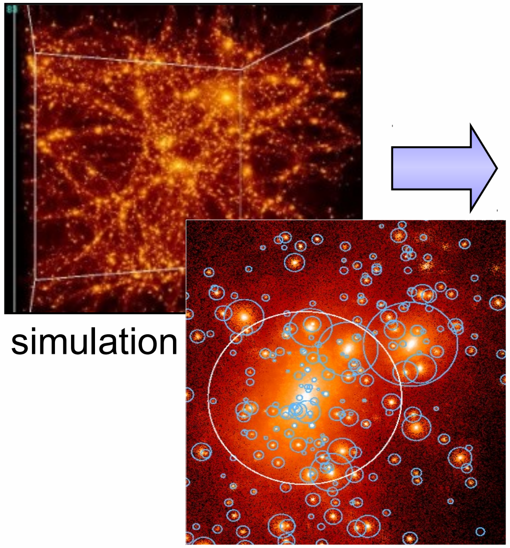
Outline

- CosmoSim introduction
- Database technology
- Data access
- Data at CosmoSim
- Usage statistics

- Discussion

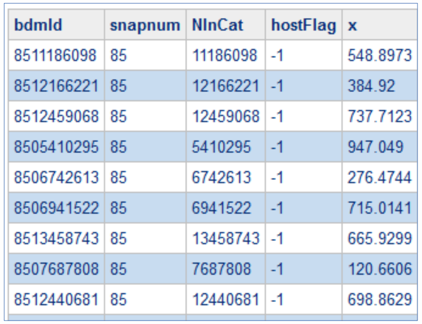
CosmoSim

- platform for publishing data from **cosmo**logical **sim**ulations
- first version: MultiDark database, collaboration with Spanish MultiDark project
- database + web query interface
- free registration, open to anyone



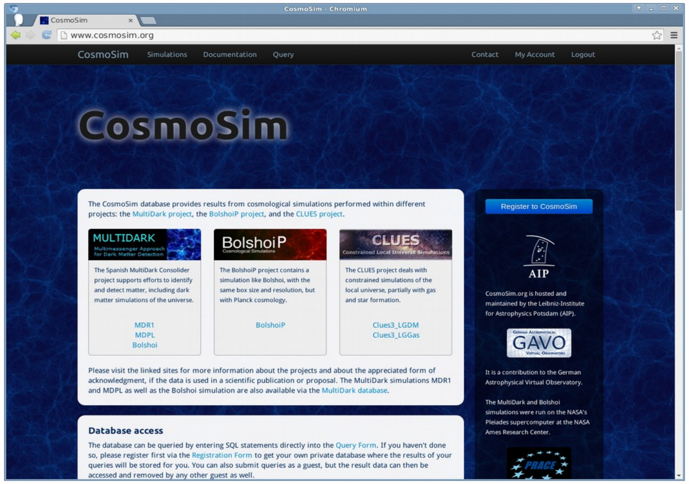
simulation

post-processing (e.g. halo finding)



bdmId	snapnum	NInCat	hostFlag	x
8511186098	85	11186098	-1	548.8973
8512166221	85	12166221	-1	384.92
8512459068	85	12459068	-1	737.7123
8505410295	85	5410295	-1	947.049
8506742613	85	6742613	-1	276.4744
8506941522	85	6941522	-1	715.0141
8513458743	85	13458743	-1	665.9299
8507687808	85	7687808	-1	120.6606
8512440681	85	12440681	-1	698.8629

data catalogue



The CosmoSim database provides results from cosmological simulations performed within different projects: the MultiDark project, the BolshoiP project, and the CLUES project.

The Spanish MultiDark Consolider project supports efforts to identify and detect matter, including dark matter simulations of the universe.

MDR1
MDPL
Bolshoi

The BolshoiP project contains a simulation like Bolshoi, with the same box size and resolution, but with Planck cosmology.

BolshoiP

The CLUES project deals with constrained simulations of the local universe, partially with gas and star formation.

Clues3_LGDM
Clues3_LGGas

Please visit the linked sites for more information about the projects and about the appreciated form of acknowledgment, if the data is used in a scientific publication or proposal. The MultiDark simulations MDR1 and MDPL as well as the Bolshoi simulation are also available via the MultiDark database.

Database access

The database can be queried by entering SQL statements directly into the Query Form. If you haven't done so, please register first via the Registration Form to get your own private database where the results of your queries will be stored for you. You can also submit queries as a guest, but the result data can then be accessed and removed by any other guest as well.

CosmoSim.org is hosted and maintained by the Leibniz-Institute for Astrophysics Potsdam (AIP).

It is a contribution to the German Astrophysical Virtual Observatory.

The MultiDark and Bolshoi simulations were run on the NASA's Pleiades supercomputer at the NASA Ames Research Center.

https://www.cosmosim.org

Why use databases?

(and not just download the files ...)

- data becomes huge, not quick to download
- let server do most of the calculations
- retrieve only subsets/results, **not** complete catalogues
- Structured Query Language: SQL, quite easy to learn
- examples: sort/filter halos, calculate mass functions, merger trees, follow mass growth of stellar disk, ...

```
SELECT * FROM MDR1.FOF
WHERE snapnum=85
ORDER BY mass DESC LIMIT 10
```

10 most massive FOF groups at z=0

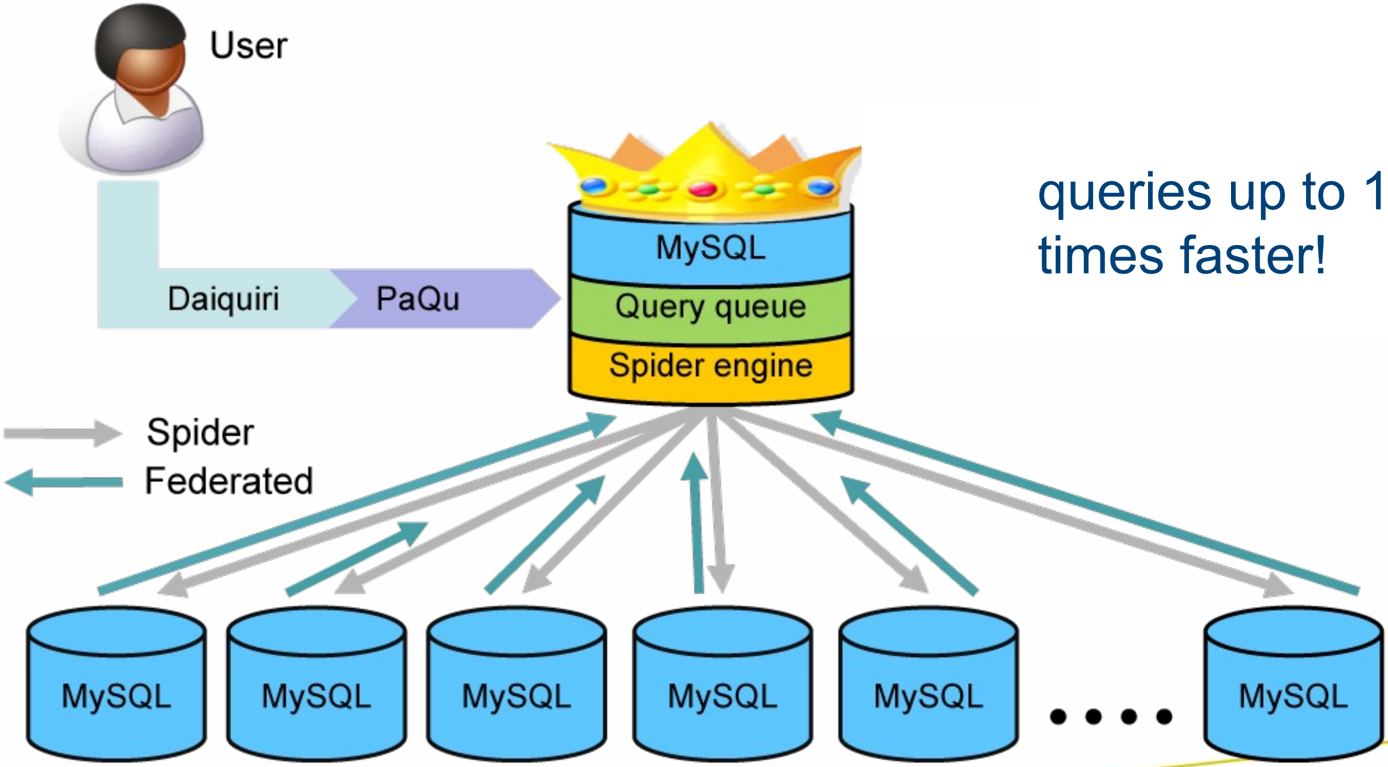
=> Just share the query, don't need to share the data!

Database server

- 10 shard nodes with distributed data, 1 head node
- MariaDB (MySQL variant)
  - MyISAM engine (no transactions => fast)
  - Spider engine for distributed queries
- open source
- own developments in E-Science group:
  - see http://github.com/aipescience
    and http://github.com/adrpar

Spider engine

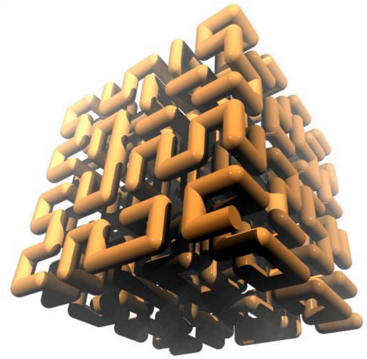
- data tables partitioned, distributed over 10 nodes engine
- PaQu reformulates queries, head node sends them to nodes
- head node collects data via federated table



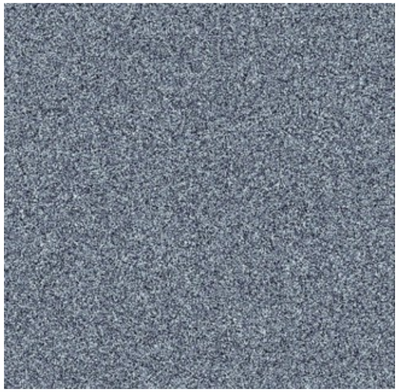
queries up to 10 times faster!

Further MySQL plugins

- C-library libhilbert
  - For creating indexes of space-filling Peano-Hilbert curve in up to 20 dimensions



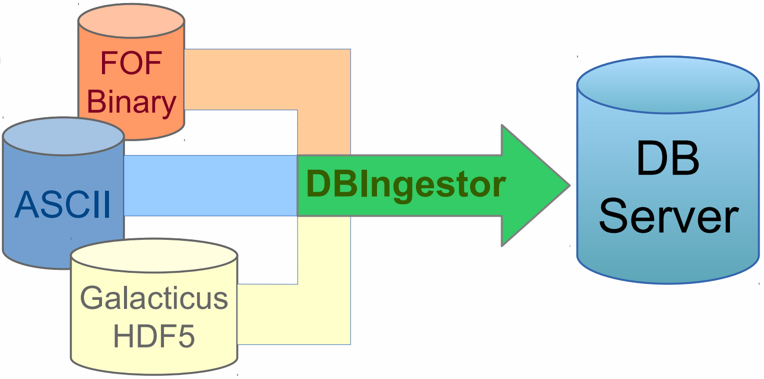
- MySQL sprng
  - Based on The Scalable Parallel Random Number Generators Library (SPRNG, www.sprng.org)
  - Implements several random number generators
  - Better random sampling for large numbers than with built-in function
  - useful for e.g. extracting a random sample of halos/galaxies



see https://github.com/adrpar/

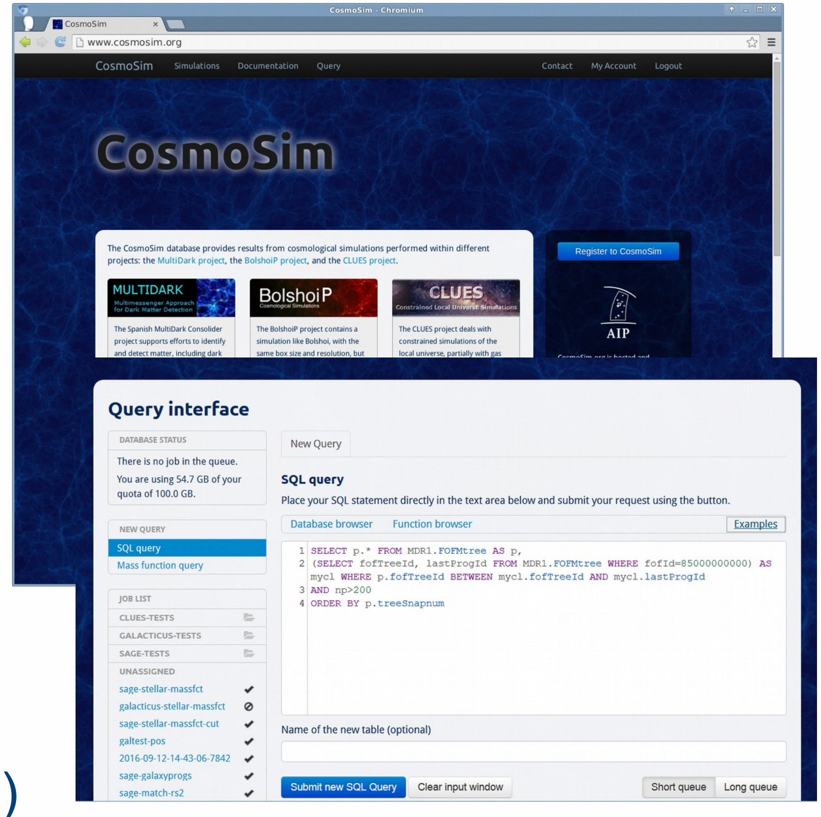
Upload: DBIngestor

- Data has variety of formats, need transformations
- DBIngestor library: one tool to load them all
  - by Adrian Partl, open source: https://github.com/aipescience/DBIngestor
  - adjustable to any database server
  - write own file readers (e.g. FofIngest, see https://github.com/kristinriebe)
  - apply converters during ingestion
    - e.g. unit conversion, type conversion (int/real), adding identifiers, grid indexes
  - apply asserters (not nan, inf, null etc.)
  - => transform and upload in one go
  - => easier to preserve the workflow for later reference



Database access: webinterface

- Daiquiri web application
  - http://escience.aip.de/daiquiri
  - developed by J. Klar and A. Partl
  - modular, highly customizable
  - using PHP, Zend-framework
  - authentication, query interface
  - wordpress integration for documentation
  - open source
  - also used for other projects (databases in Madrid, Gaia at AIP)
  - use SAMP for sending results to VO clients (Topcat) directly from webpage



The CosmoSim database provides results from cosmological simulations performed within different projects: the MultiDark project, the BolshoiP project, and the CLUES project.

The Spanish MultiDark Consolider project supports efforts to identify and detect matter, including dark

The BolshoiP project contains a simulation like Bolshoi, with the same box size and resolution, but

The CLUES project deals with constrained simulations of the local universe, partially with gas

There is no job in the queue.

You are using 54.7 GB of your quota of 100.0 GB.

sage-stellar-massfct ✓
galactitus-stellar-massfct ⊘
sage-stellar-massfct-cut ✓
galtest-pos ✓
2016-09-12-14-43-06-7842 ✓
sage-galaxyprogs ✓
sage-match-rs2 ✓

SQL query

Place your SQL statement directly in the text area below and submit your request using the button.

```
SELECT p.* FROM MDR1.FOFMtree AS p,
(SELECT fofTreeId, lastProgId FROM MDR1.FOFMtree WHERE fofId=85000000000) AS mycl WHERE p.fofTreeId BETWEEN mycl.fofTreeId AND mycl.lastProgId
AND np>200
ORDER BY p.treeSnapnum
```

Name of the new table (optional)

Database access

- UWS
  - universal worker service, virtual observatory standard
  - interfaces to create, execute, abort or delete query jobs
  - write scripts for submitting many jobs at once
- different tools available:
  - **httpie/curl/wget:**

    ```
    http --auth <username>:<password> --print b GET
    https://www.cosmosim.org/uws/query
    ```

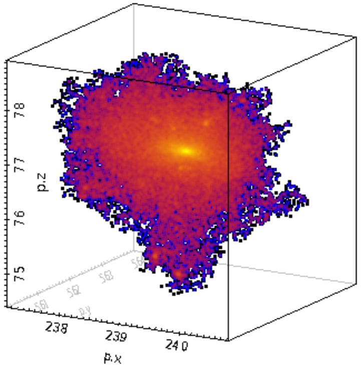
  - **uws-client** (https://github.com/aipescience/uws-client):
    - python client, supports latest features

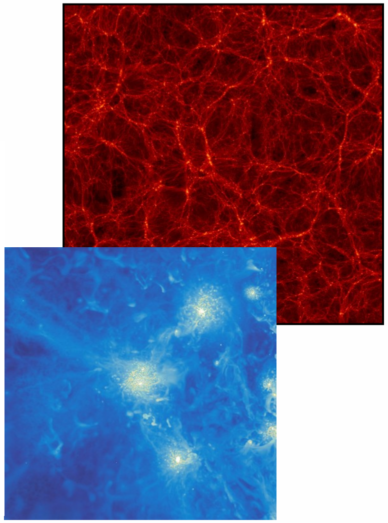
    ```
    uws --host https://www.cosmosim.org/uws/query --user
    <username> --password <password> list
    ```

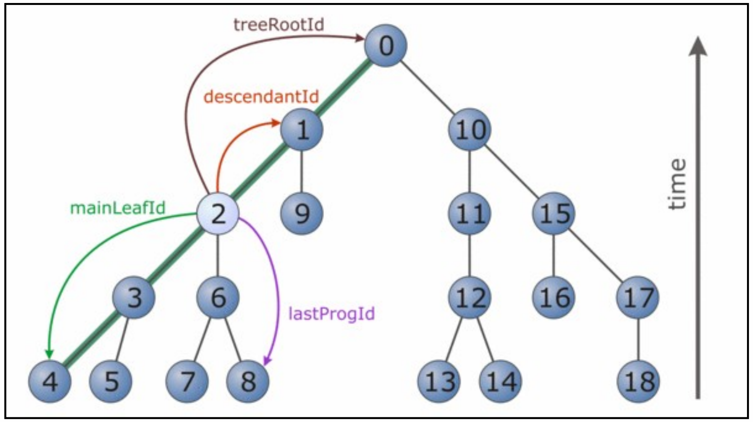
  - **cosmosim**-package for „**astroquery**“ by Austen Groener:
    - https://github.com/astropy/astroquery/cosmosim
      (Beware: needs to be updated!)

Data at CosmoSim

- about 40 TB data, ~ 350 billion rows
- 10 simulations
- box sizes: 64 Mpc/h -- 4 Gpc/h
- data products:
  - simulation particles
  - density fields
  - dark matter halo catalogues
  - dark matter profiles
  - merger trees





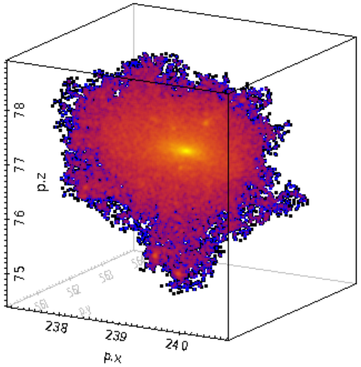


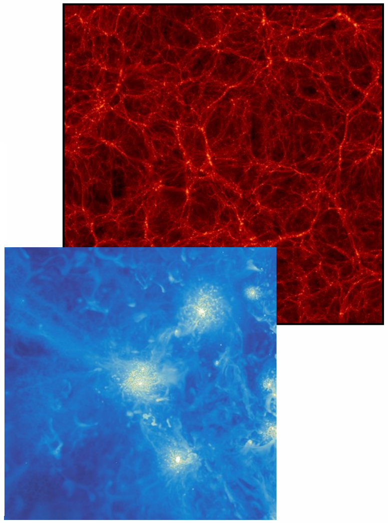
depth first order for fast retrieval of merger trees

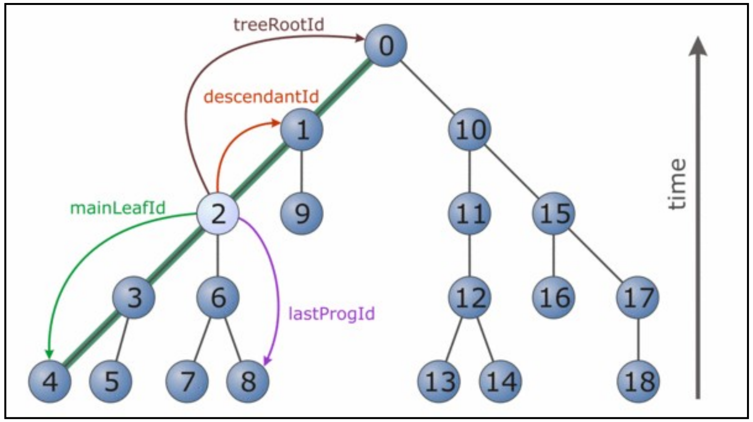
Data at CosmoSim

- about 40 TB data, ~ 350 billion rows
- 10 simulations
- box sizes: 64 Mpc/h -- 4 Gpc/h
- data products:
  - simulation particles
  - density fields
  - dark matter halo catalogues
  - dark matter profiles
  - merger trees

+ galaxy data

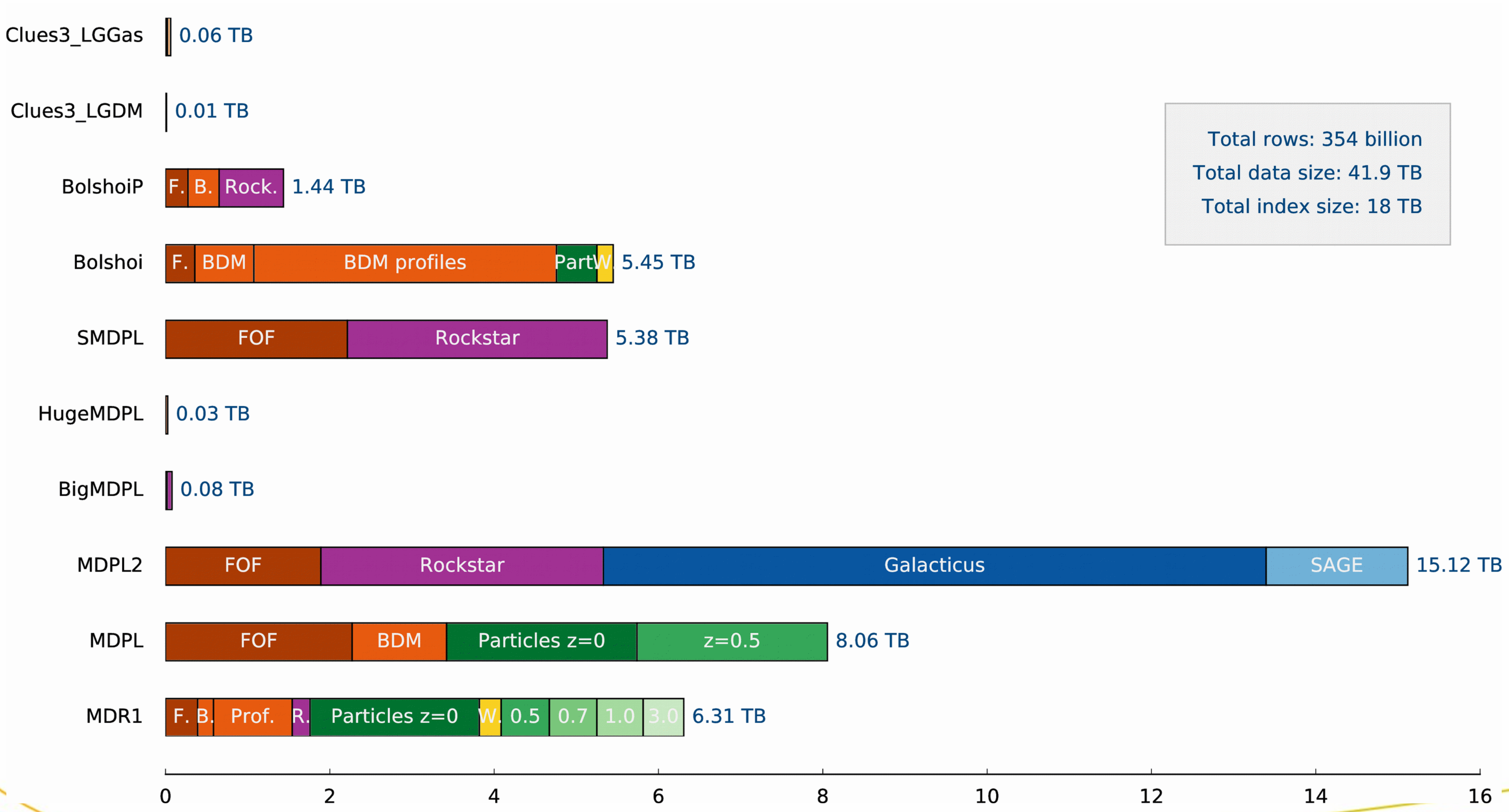






depth first order for fast retrieval of merger trees

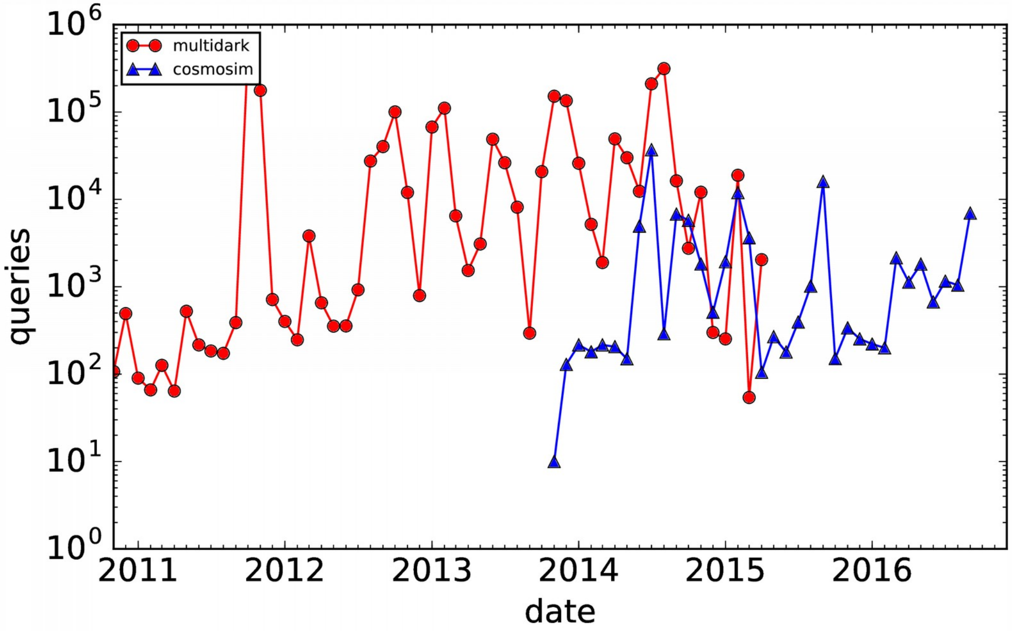
Data at CosmoSim



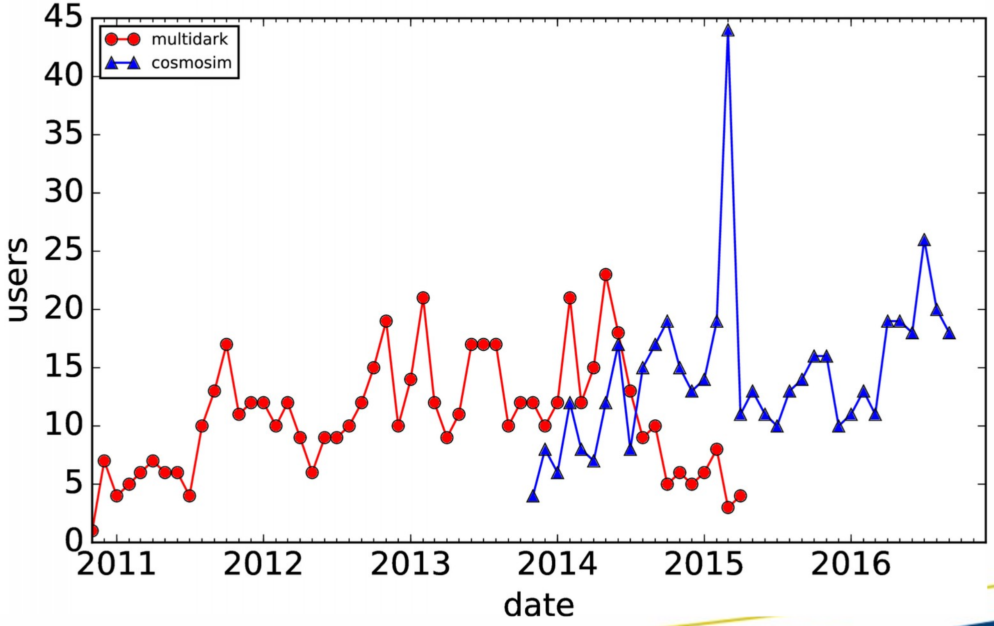
Total rows: 354 billion
Total data size: 41.9 TB
Total index size: 18 TB

User statistics

- Number of queries per month



- Number of (unique) users per month



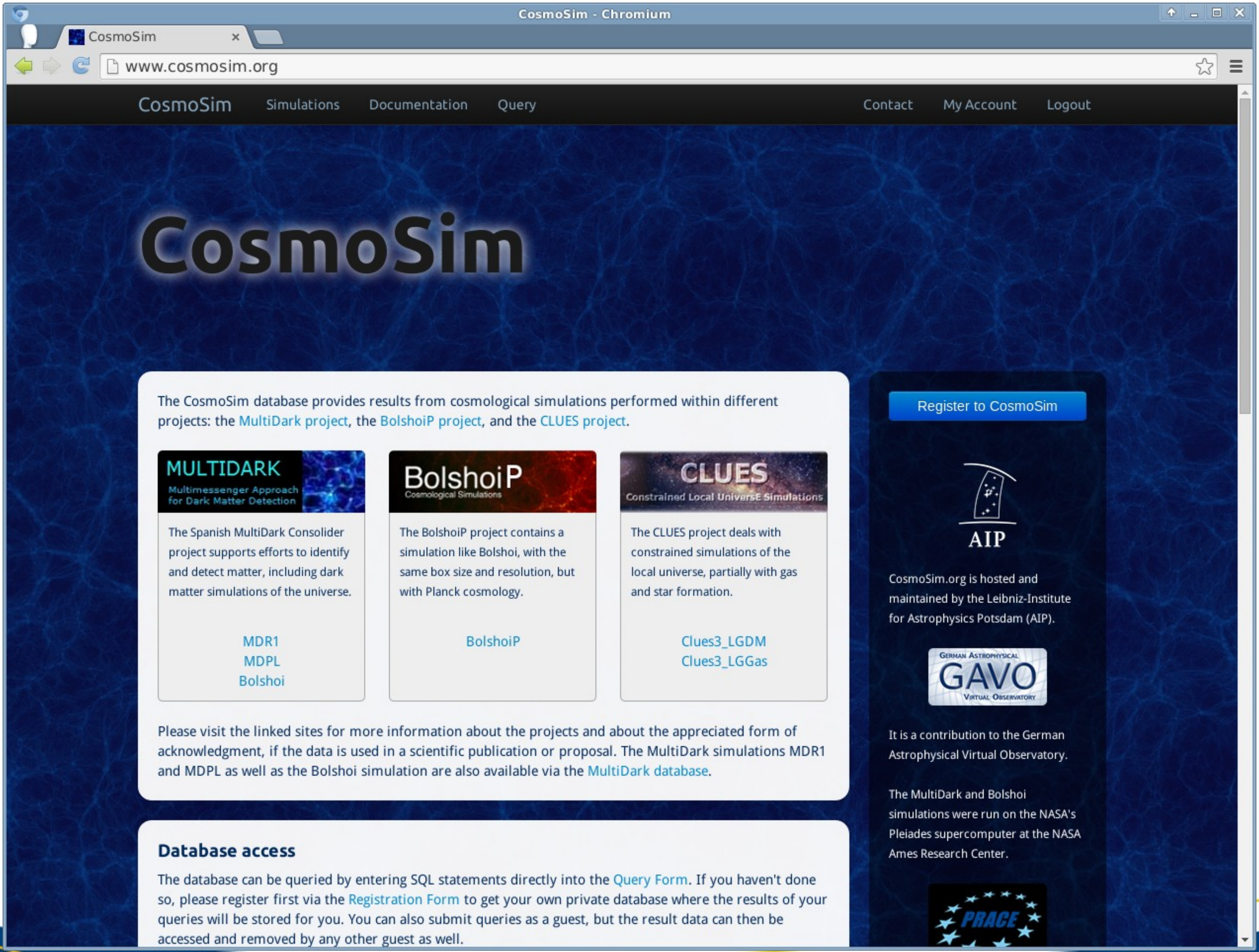
400 users registered, 178 with at least 1 query

Most wanted data on CosmoSim

- MDR1: 640,000 queries; 154 users
- Bolshoi: 34,000 queries; 89 users
- MDPL2: 6,500 queries; 34 users

- Particles: 79,000 queries; 82 users
- FOFParticles: 37,500 queries; 43 users
- FOF: 53,000 queries; 107 users
- BDM: 23,000 queries; 86 users

Demo: querying data from CosmoSim



The CosmoSim database provides results from cosmological simulations performed within different projects: the MultiDark project, the BolshoiP project, and the CLUES project.

The Spanish MultiDark Consolider project supports efforts to identify and detect matter, including dark matter simulations of the universe.

MDR1
MDPL
Bolshoi

The BolshoiP project contains a simulation like Bolshoi, with the same box size and resolution, but with Planck cosmology.

BolshoiP

The CLUES project deals with constrained simulations of the local universe, partially with gas and star formation.

Clues3_LGDM
Clues3_LGGas

Please visit the linked sites for more information about the projects and about the appreciated form of acknowledgment, if the data is used in a scientific publication or proposal. The MultiDark simulations MDR1 and MDPL as well as the Bolshoi simulation are also available via the MultiDark database.

Database access

The database can be queried by entering SQL statements directly into the Query Form. If you haven't done so, please register first via the Registration Form to get your own private database where the results of your queries will be stored for you. You can also submit queries as a guest, but the result data can then be accessed and removed by any other guest as well.

CosmoSim.org is hosted and maintained by the Leibniz-Institute for Astrophysics Potsdam (AIP).

It is a contribution to the German Astrophysical Virtual Observatory.

The MultiDark and Bolshoi simulations were run on the NASA's Pleiades supercomputer at the NASA Ames Research Center.

Discussion

- Galaxy data:
  - When will final SAG version be available?
  - Which other formats? (Galform?)
  - Time plan for publishing galaxy data? Papers?
  - Generate DOIs for citations?
- Which simulations next?
  - SMDPL?
  - Time frame?
- New data:
  - need sample data with description + sample read routine
  - estimated data volume
  - time scale for intended publication
  - updates on database are slow (depending on data volume ...)
  - => upload data already into finalized format