The QUBRICS database for machine learning: architecture and performance.

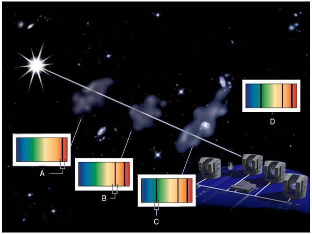
Giorgio Calderone (INAF-OATs)

QUBRICS collaboration:
Konstantina Boutsia, Stefano Cristiani, Guido Cupani,
Andrea Grazian, Francesco Guarneri, Luciano Nicastro, Matteo Porru, ...

© KWON, O CHUL

The QUBRICS Project

QUasars as BRIght beacons for Cosmology in the Southern hemisphere

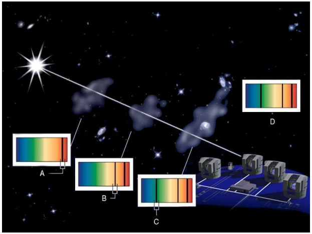


Purpose:

- Collect photometric datasets in the Southern Hemisphere;
- Use machine learning to select new, bright, high-z ($z > 2.5$) QSOs candidates;
- Spectroscopic follow-up, confirm classification and redshift;
- ⇒ exploit acquired knowledge!

The QUBRICS Project

QUasars as BRIght beacons for Cosmology in the Southern hemisphere



Purpose:

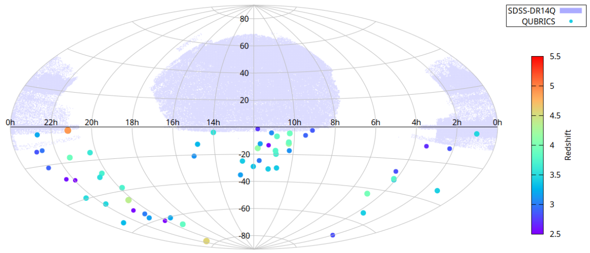
- Collect photometric datasets in the Southern Hemisphere;
- Use machine learning to select new, bright, high-z ($z > 2.5$) QSOs candidates;
- Spectroscopic follow-up, confirm classification and redshift;
- $\Rightarrow$ exploit acquired knowledge!

Papers:

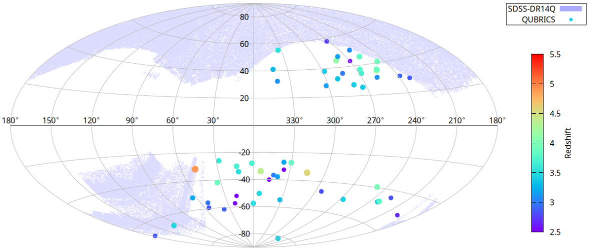
Calderone+19, Boutsia+20, Boutsia+21, Guarneri+21, Cupani+21, Grazian+21, Guarneri+22, Cristiani+23, Grazian+23, Calderone+24, Grazian+24,
More in preparation...

QUBRICS QSOs ($z > 2.5$) in 2019

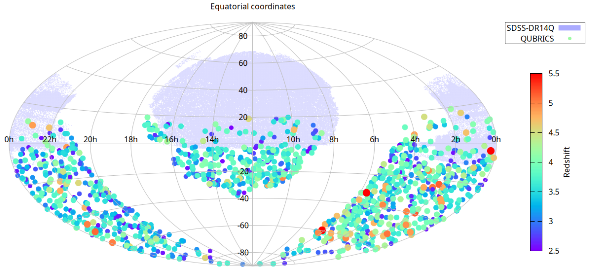
Equatorial coordinates

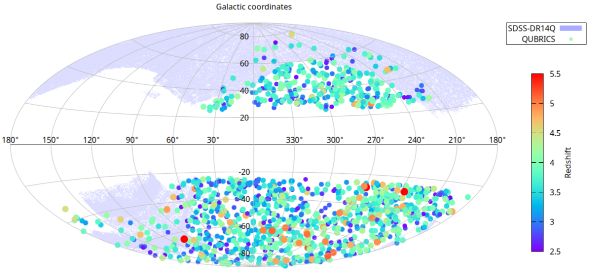


Galactic coordinates

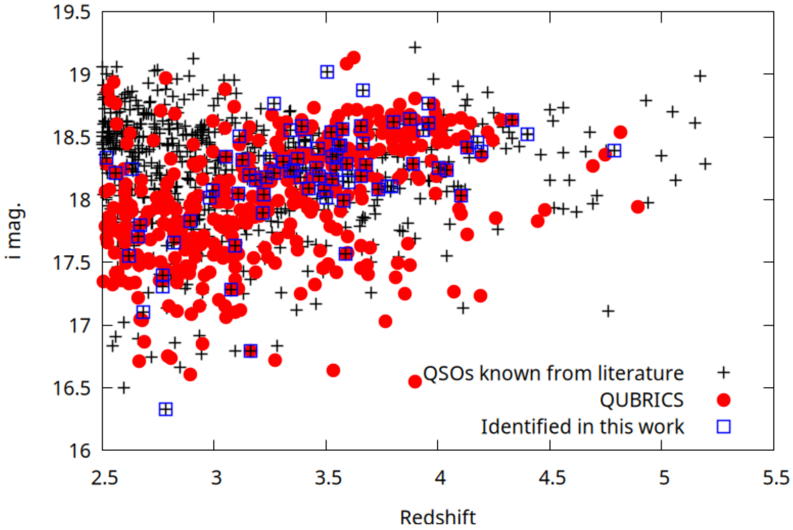


QUBRICS QSOs ($z > 2.5$) in 2025


Equatorial coordinates


Galactic coordinates

QUBRICS QSOs vs mag. in *i* band

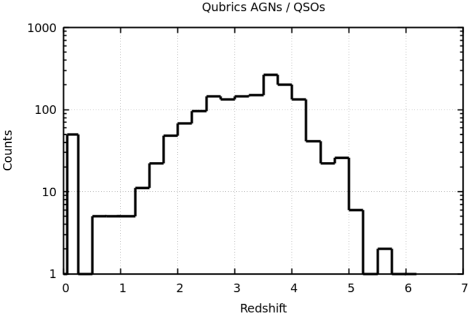


Calderone et al., 2024

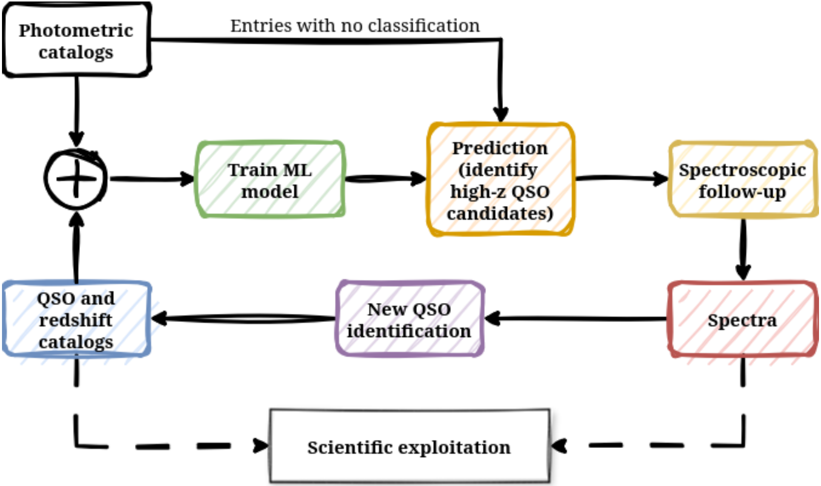
QUBRICS statistics in 2025

Observations:

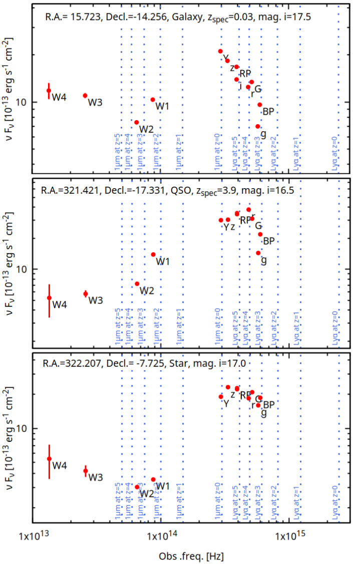
- Candidates observed: 2019;
  - good quality: 1764;
    - Stars: 76;
    - Galaxies: 38;
  - bad quality: 255.
- QSOs: 1585 (93%);
  - z>2.0: 1438 (84%);
  - z>2.5: 1274 (75%);
  - z>3.0: 992 (58%);
  - z>4.0: 228 (13%)
- Max. redshift: 5.768.



QUBRICS self-feeding



How data looks like?

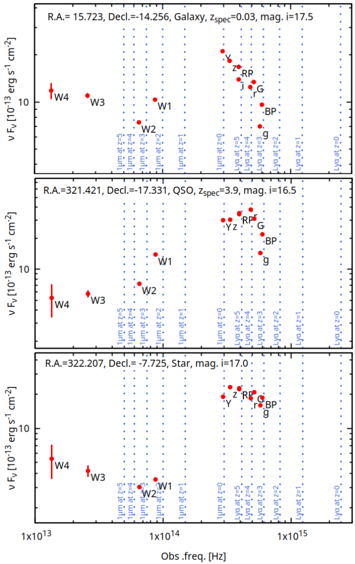


R.A.= 15.723, Decl.=-14.256, Galaxy, z_{spec}=0.03, mag. i=17.5

R.A.=321.421, Decl.=-17.331, QSO, z_{spec}=3.9, mag. i=16.5

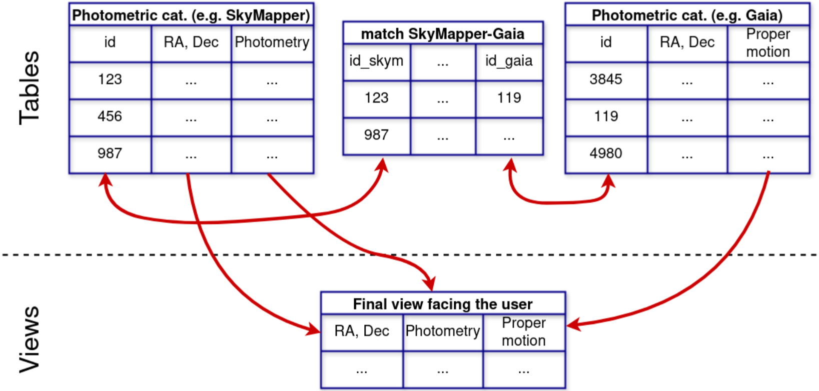
R.A.=322.207, Decl.= -7.725, Star, mag. i=17.0

How data looks like?



- Disk usage ∼ 4TB;
- Single workstation with 8 CPU, 64GB RAM;
- MariaDB database;
- Less than 10 persons involved (+occasional contributors)
- ⇒ **“Small data” project**.

Relational DB, inner joins



Photometric cat. (e.g. SkyMapper)		
id	RA, Dec	Photometry
123	...	...
456	...	...
987	...	...

match SkyMapper-Gaia		
id_skym	...	id_gaia
123	...	119
987	...	...

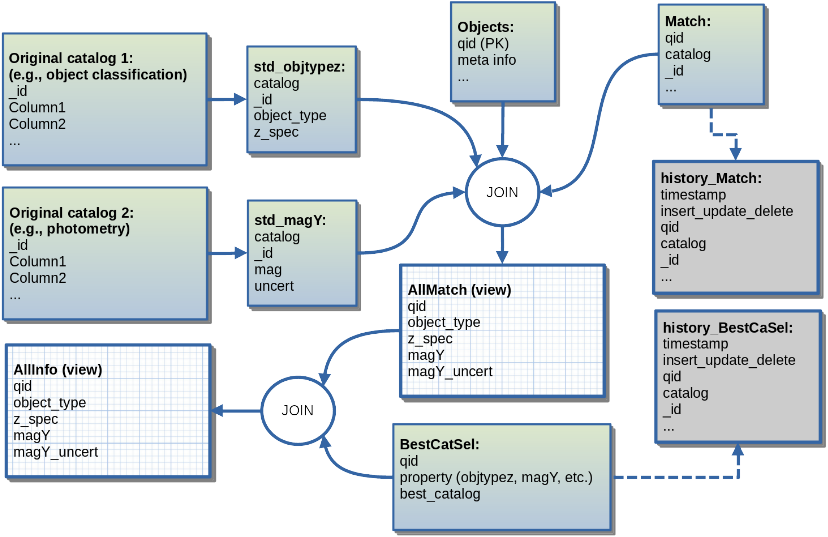
Photometric cat. (e.g. Gaia)		
id	RA, Dec	Proper motion
3845	...	...
119	...	...
4980	...	...

Final view facing the user		
RA, Dec	Photometry	Proper motion
...	...	...

- Original catalogs are not modified;
- “Matching” tables contain matching entries among two tables;

- User make queries on a “view”;
- **It is always possible to trace back entries to their original catalogs;**

QUBRICS Database



Original catalog 1:
(e.g., object classification)
_id
Column1
Column2
...

std_objtypez:
catalog
_id
object_type
z_spec

Objects:
qid (PK)
meta info
...

Match:
qid
catalog
_id
...

Original catalog 2:
(e.g., photometry)
_id
Column1
Column2
...

std_magY:
catalog
_id
mag
uncert

history_Match:
timestamp
insert_update_delete
qid
catalog
_id
...

AllMatch (view)
qid
object_type
z_spec
magY
magY_uncert

history_BestCaSel:
timestamp
insert_update_delete
qid
catalog
_id
...

AllInfo (view)
qid
object_type
z_spec
magY
magY_uncert

BestCatSel:
qid
property (objtypez, magY, etc.)
best_catalog

QUBRICS Database

DB Content:

- Photometric: Gaia DR3, PanSTARRS1 DR2, DES DR2, SkyMapper DR4, AllWise and CatWISE, SDSS DR16Q, etc. ;
- Several QSO and inactive galaxies catalogs;
- Stars are identified using Gaia's parallax and proper motion measurements;

QUBRICS Database

DB Content:

- Photometric: Gaia DR3, PanSTARRS1 DR2, DES DR2, SkyMapper DR4, AllWise and CatWISE, SDSS DR16Q, etc. ;
- Several QSO and inactive galaxies catalogs;
- Stars are identified using Gaia's parallax and proper motion measurements;

Performance

- Select data for a single QSO: $\sim$ 1 ms;
- Query entire table containing coordinates, classifications and redshifts (8×10^5 rows) of known QSOs: $\sim$ 14 s;
- Query on photometric catalogs is significantly slower ($\sim 10^8$ rows): pre-matched and stored as separate table;

QUBRICS Database

DB Content:

- Photometric: Gaia DR3, PanSTARRS1 DR2, DES DR2, SkyMapper DR4, AllWise and CatWISE, SDSS DR16Q, etc. ;
- Several QSO and inactive galaxies catalogs;
- Stars are identified using Gaia's parallax and proper motion measurements;

Performance

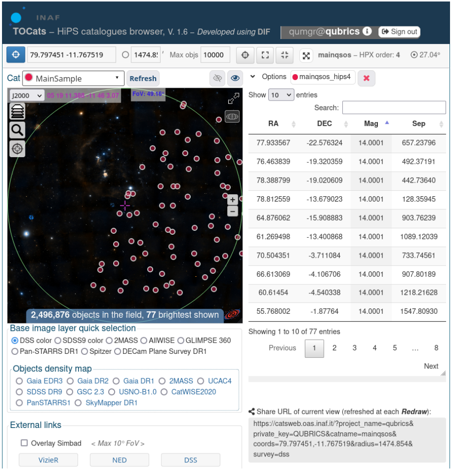
- Select data for a single QSO: $\sim$ 1 ms;
- Query entire table containing coordinates, classifications and redshifts (8×10^5 rows) of known QSOs: $\sim$ 14 s;
- Query on photometric catalogs is significantly slower ($\sim 10^8$ rows): pre-matched and stored as separate table;

Features:

- All tables have indices on their primary keys, as well as on coordinates: access time do not depends on table size;
- Main table modification histories are recorded;
- DB coherence ensured by triggers.

QUBRICS on TOCats

- Easy visualization of catalogs on the sky;
- Uses multi-depth indexing to **quickly** access catalogs with $\sim 10^8$ entries;
- Quick access to online services and private repository of spectra;
- $\Rightarrow$ talk by L. Nicastro.



TOCats – HiPS catalogues browser, V. 1.6 – *Developed using* DIF

qumgr@**qubrics** Sign out

79.797451 -11.767519 ○ 1474.8' Max objs 10000 **mainqsos** – HPX order: 4 27.04°

Cat ● MainSample Refresh

Options ● mainqsos_hips4

2,496,876 objects in the field, **77** brightest shown

Show 10 entries

Search:

RA	DEC	Mag	Sep
77.933567	-22.576324	14.0001	657.23796
76.463839	-19.320359	14.0001	492.37191
78.388799	-19.020609	14.0001	442.73640
78.812559	-13.679023	14.0001	128.35945
64.876062	-15.908883	14.0001	903.76239
61.269498	-13.400868	14.0001	1089.12039
70.504351	-3.711084	14.0001	733.74561
66.613069	-4.106706	14.0001	907.80189
60.61454	-4.540338	14.0001	1218.21628
55.768002	-1.87764	14.0001	1547.80930

Showing 1 to 10 of 77 entries

Previous 1 2 3 4 5 … 8 Next

Base image layer quick selection

● DSS color ○ SDSS9 color ○ 2MASS ○ AllWISE ○ GLIMPSE 360 ○ Pan-STARRS DR1 ○ Spitzer ○ DECam Plane Survey DR1

Objects density map

○ Gaia EDR3 ○ Gaia DR2 ○ Gaia DR1 ○ 2MASS ○ UCAC4 ○ SDSS DR9 ○ GSC 2.3 ○ USNO-B1.0 ○ CatWISE2020 ○ PanSTARRS1 ○ SkyMapper DR1

External links

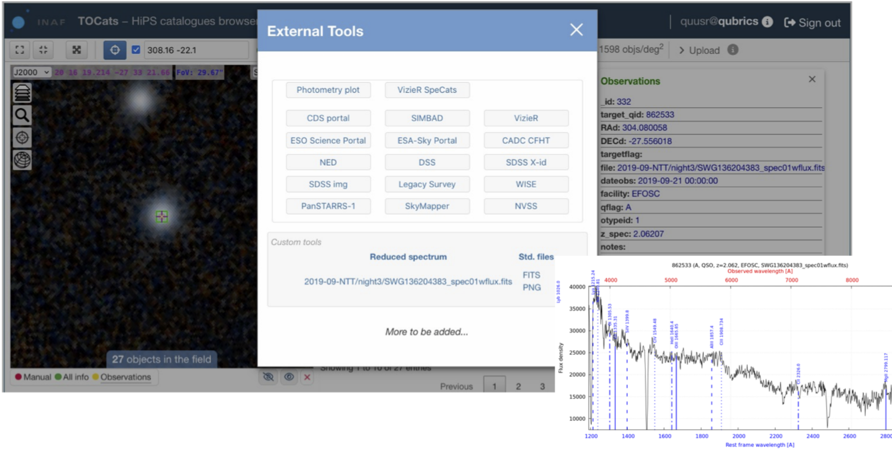
☐ Overlay Simbad *< Max 10° FoV >*

VizieR NED DSS

Share URL of current view (refreshed at each ***Redraw***):

https://catsweb.oas.inaf.it/?project_name=qubrics&private_key=QUBRICS&catname=mainqsos&coords=79.797451,-11.767519&radius=1474.854&survey=dss

QUBRICS on TOCats



INAF TOCats – HiPS catalogues browser

quusr@qubrics ⓘ Sign out

308.16 -22.1

J2000 20 16 19.214 -27 33 21.66 FoV: 29.67"

1598 objs/deg^2

› Upload ⓘ

External Tools

Photometry plot | VizieR SpeCats

CDS portal | SIMBAD | VizieR

ESO Science Portal | ESA-Sky Portal | CADC CFHT

NED | DSS | SDSS X-id

SDSS img | Legacy Survey | WISE

PanSTARRS-1 | SkyMapper | NVSS

Custom tools

Reduced spectrum	Std. files
2019-09-NTT/night3/SWG136204383_spec01wflux.fits	FITS PNG

More to be added...

27 objects in the field

● Manual ● All info ● Observations

Showing 1 to 10 of 27 entries

Previous 1 2 3

Observations

_id: 332
target_qid: 862533
RAd: 304.080058
DECd: -27.556018
targetflag:
file: 2019-09-NTT/night3/SWG136204383_spec01wflux.fits
dateobs: 2019-09-21 00:00:00
facility: EFOSC
qflag: A
otypeid: 1
z_spec: 2.06207
notes:

Summary

- The QUBRICS project already discovered more than 1000 QSO at $z > 2.5$ (and more than 200 at $z > 4$) at Y $\lesssim$ 19.5;
  - Scientific exploitation: luminosity function, cosmological re-ionization, Sandage test, etc.;
- Several machine learning methods adopted: CCA, PRF, XGBoost;
- Dedicated method to deal with severely imbalanced datasets (Calderone+24);

- QUBRICS database is a key part in the project;
- Storage for machine learning project *may* be smaller than for instrumentation...
  - ...and typically has a clear "structure" (feature columns) ⇒ relational DB is the perfect tool!
- By using pre-matched tables (e.g. on coordinates), indices, views, etc. we built a highly responsive DB able to support all project activities;
- TOCats is the perfect companion for our DB, to quickly visualize literature data and prioritize observations;

Summary

- The QUBRICS project already discovered more than 1000 QSO at $z > 2.5$ (and more than 200 at $z > 4$) at Y $\lesssim$ 19.5;
  - Scientific exploitation: luminosity function, cosmological re-ionization, Sandage test, etc.;
- Several machine learning methods adopted: CCA, PRF, XGBoost;
- Dedicated method to deal with severely imbalanced datasets (Calderone+24);

- QUBRICS database is a key part in the project;
- Storage for machine learning project *may* be smaller than for instrumentation...
  - ...and typically has a clear “structure” (feature columns) $\Rightarrow$ relational DB is the perfect tool!
- By using pre-matched tables (e.g. on coordinates), indices, views, etc. we built a highly responsive DB able to support all project activities;
- TOCats is the perfect companion for our DB, to quickly visualize literature data and prioritize observations;

Summary

Future challenges:

- Add further photometric data to the NIR (e.g. VISTA, Euclid, etc.);
- Optimize ML selection;
- Probe redshifts up to $\sim$ 5.5 - 6, with completeness $\gtrsim$ 90%;

QUBRICS activities range from technical topics, such as DB management and machine learning, to the science exploitation.

If you're interested in learning/collaborating, feel free to reach us. Students are very welcome! ;-)