Outils de cross-correlation : accès aux catalogues et cross-identifications

F.-X. Pineau[1] et l'équipe du CDS

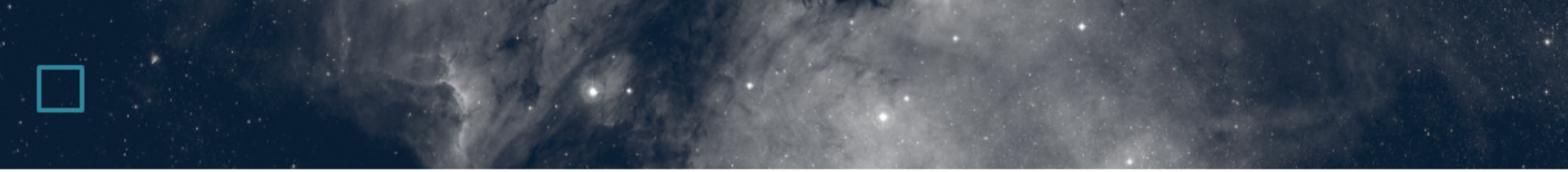
[1] CDS, Observatoire Astronomique de Strasbourg

Transient Sky 2020 National Workshop - 2nd Edition
Montpellier, 4-6th June, 2018

Université de Strasbourg | Observatoire astronomique de Strasbourg | ObAS | CDS Centre de Données astronomiques de Strasbourg

Plan

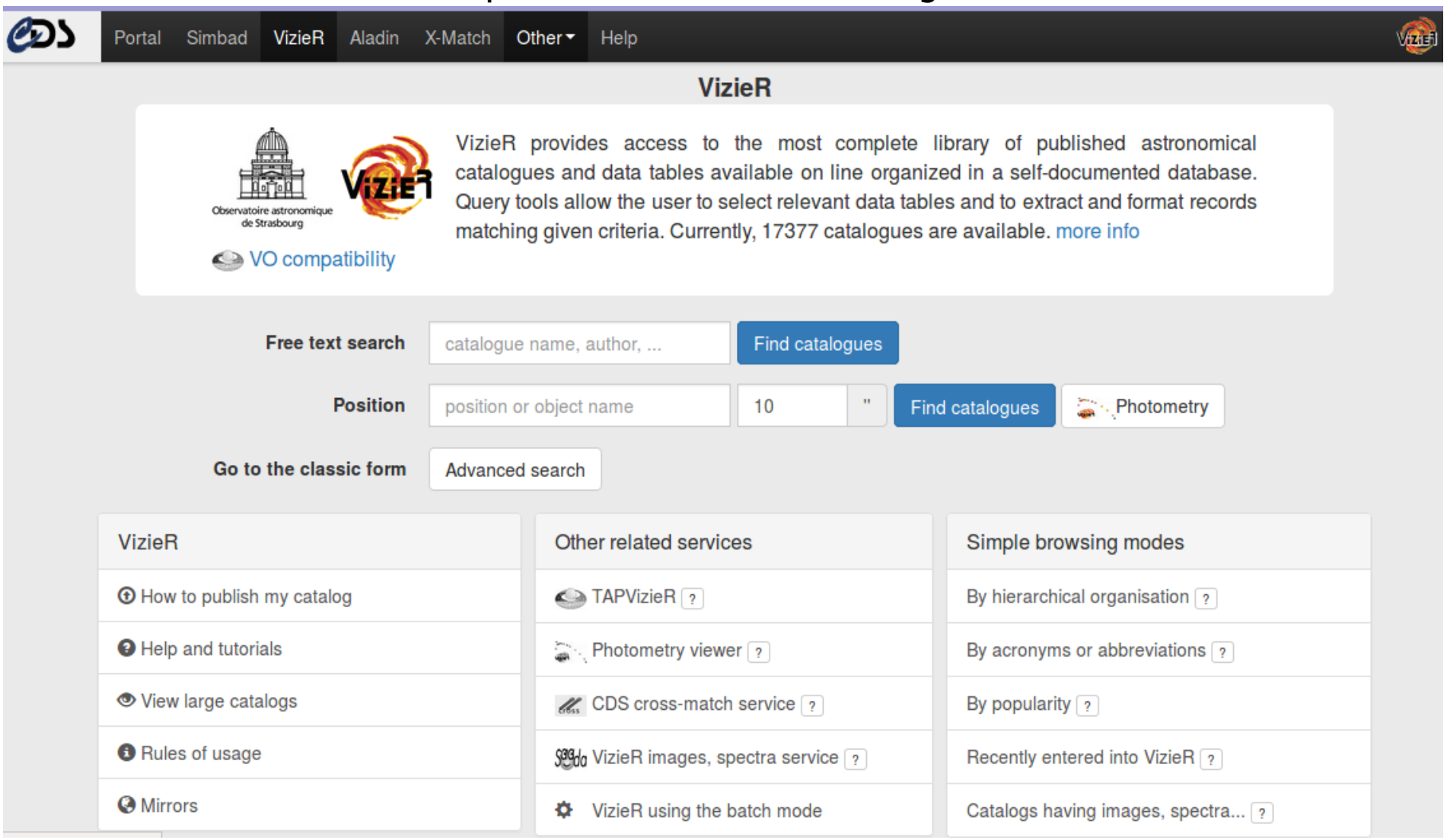
- Accès aux catalogues du CDS
  - tables normales, grandes tables
- Cross-match simple
  - service de xmatch du CDS
- Cross-match complexe
  - outil ARCHES
- Prise en compte de paramètres photométriques
  - estimation de vraissemblances photométriques par noyau
- Développements exploratoires:
  - Index spatio-temporel basé sur HEALPix
  - Test de Apache Spark



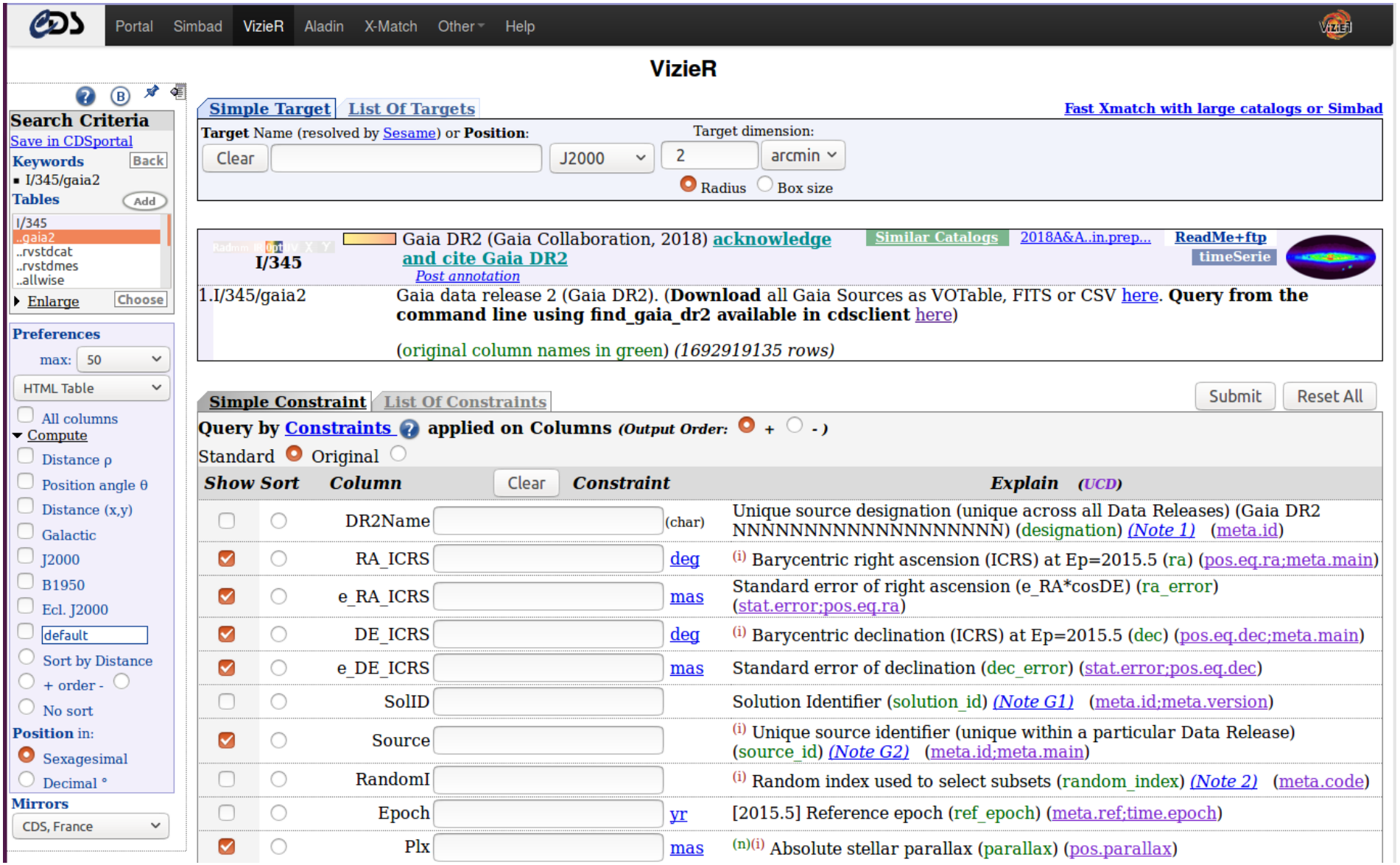
Accès aux catalogues

VizieR: interface web

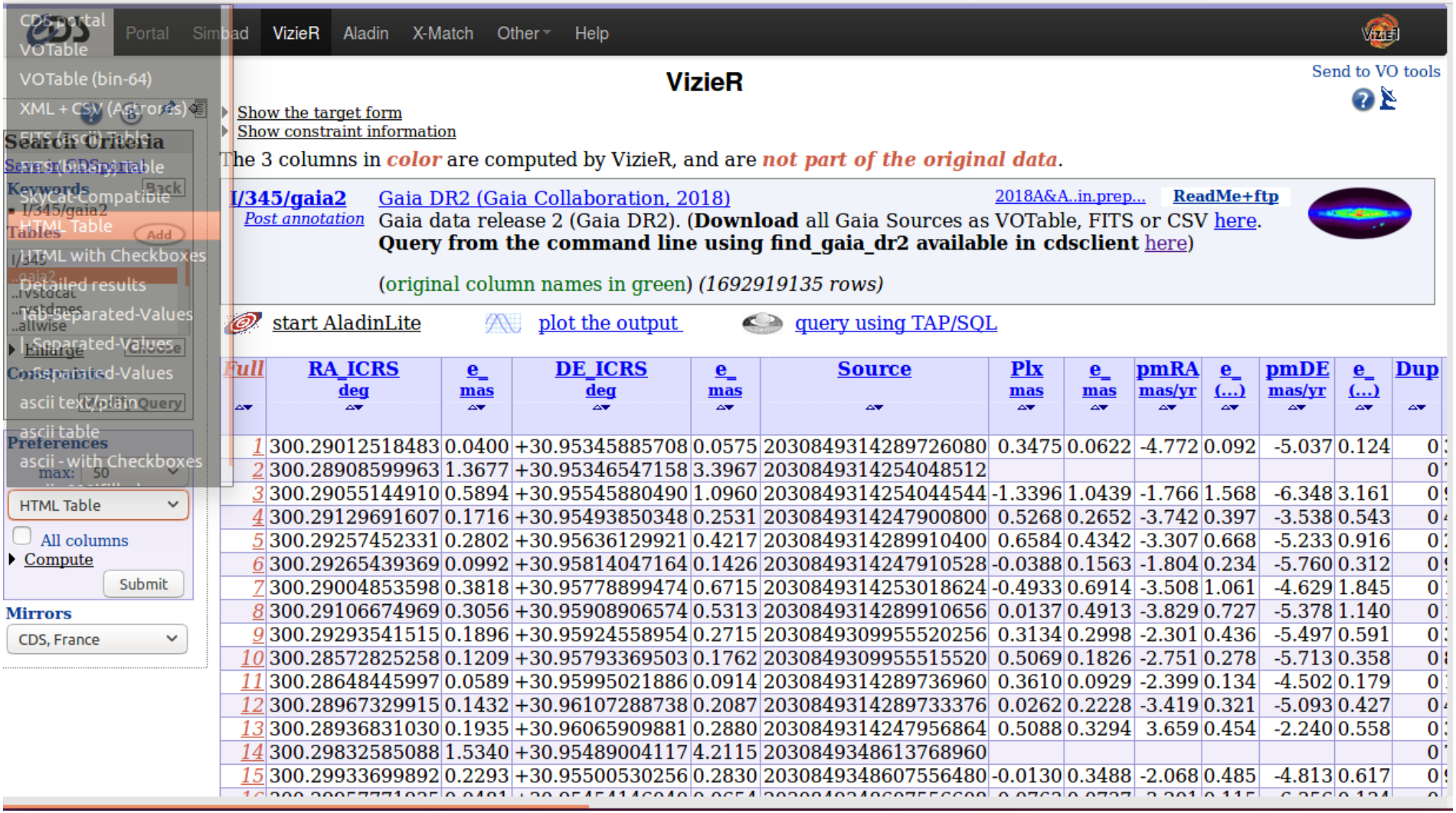
http://vizier.u-strasbg.fr/



VizieR: interface web



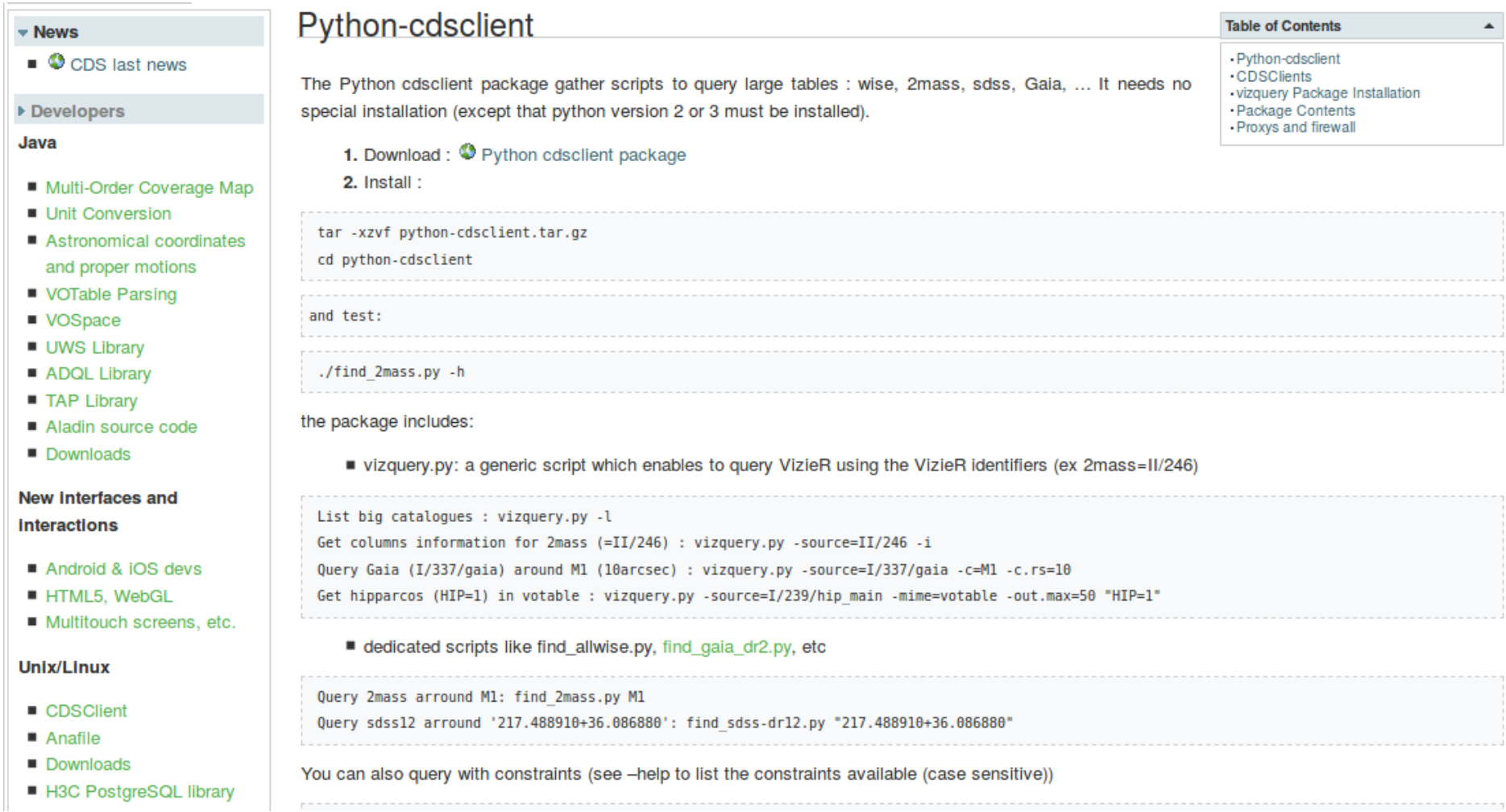
VizieR: interface web



Full	RA_ICRS deg	e_ mas	DE_ICRS deg	e_ mas	Source	Plx mas	e_ mas	pmRA mas/yr	e_ (...)	pmDE mas/yr	e_ (...)	Dup
1	300.29012518483	0.0400	+30.95345885708	0.0575	2030849314289726080	0.3475	0.0622	-4.772	0.092	-5.037	0.124	0
2	300.28908599963	1.3677	+30.95346547158	3.3967	2030849314254048512							0
3	300.29055144910	0.5894	+30.95545880490	1.0960	2030849314254044544	-1.3396	1.0439	-1.766	1.568	-6.348	3.161	0
4	300.29129691607	0.1716	+30.95493850348	0.2531	2030849314247900800	0.5268	0.2652	-3.742	0.397	-3.538	0.543	0
5	300.29257452331	0.2802	+30.95636129921	0.4217	2030849314289910400	0.6584	0.4342	-3.307	0.668	-5.233	0.916	0
6	300.29265439369	0.0992	+30.95814047164	0.1426	2030849314247910528	-0.0388	0.1563	-1.804	0.234	-5.760	0.312	0
7	300.29004853598	0.3818	+30.95778899474	0.6715	2030849314253018624	-0.4933	0.6914	-3.508	1.061	-4.629	1.845	0
8	300.29106674969	0.3056	+30.95908906574	0.5313	2030849314289910656	0.0137	0.4913	-3.829	0.727	-5.378	1.140	0
9	300.29293541515	0.1896	+30.95924558954	0.2715	2030849309955520256	0.3134	0.2998	-2.301	0.436	-5.497	0.591	0
10	300.28572825258	0.1209	+30.95793369503	0.1762	2030849309955515520	0.5069	0.1826	-2.751	0.278	-5.713	0.358	0
11	300.28648445997	0.0589	+30.95995021886	0.0914	2030849314289736960	0.3610	0.0929	-2.399	0.134	-4.502	0.179	0
12	300.28967329915	0.1432	+30.96107288738	0.2087	2030849314289733376	0.0262	0.2228	-3.419	0.321	-5.093	0.427	0
13	300.28936831030	0.1935	+30.96065909881	0.2880	2030849314247956864	0.5088	0.3294	3.659	0.454	-2.240	0.558	0
14	300.29832585088	1.5340	+30.95489004117	4.2115	2030849348613768960							0
15	300.29933699892	0.2293	+30.95500530256	0.2830	2030849348607556480	-0.0130	0.3488	-2.068	0.485	-4.813	0.617	0

VizieR: accès scriptable

http://cds.u-strasbg.fr/resources/doku.php?id=cdsclient



News

- CDS last news

Developers

Java

- Multi-Order Coverage Map
- Unit Conversion
- Astronomical coordinates and proper motions
- VOTable Parsing
- VOSpace
- UWS Library
- ADQL Library
- TAP Library
- Aladin source code
- Downloads

New Interfaces and Interactions

- Android & iOS devs
- HTML5, WebGL
- Multitouch screens, etc.

Unix/Linux

- CDSClient
- Anafile
- Downloads
- H3C PostgreSQL library

Python-cdsclient

Table of Contents

- Python-cdsclient
- CDSClients
- vizquery Package Installation
- Package Contents
- Proxys and firewall

The Python cdsclient package gather scripts to query large tables : wise, 2mass, sdss, Gaia, ... It needs no special installation (except that python version 2 or 3 must be installed).

1. Download : Python cdsclient package
2. Install :

```
tar -xzvf python-cdsclient.tar.gz
cd python-cdsclient
```

```
and test:
```

```
./find_2mass.py -h
```

the package includes:

- vizquery.py: a generic script which enables to query VizieR using the VizieR identifiers (ex 2mass=II/246)

```
List big catalogues : vizquery.py -l
Get columns information for 2mass (=II/246) : vizquery.py -source=II/246 -i
Query Gaia (I/337/gaia) around M1 (10arcsec) : vizquery.py -source=I/337/gaia -c=M1 -c.rs=10
Get hipparcos (HIP=1) in votable : vizquery.py -source=I/239/hip_main -mime=votable -out.max=50 "HIP=1"
```

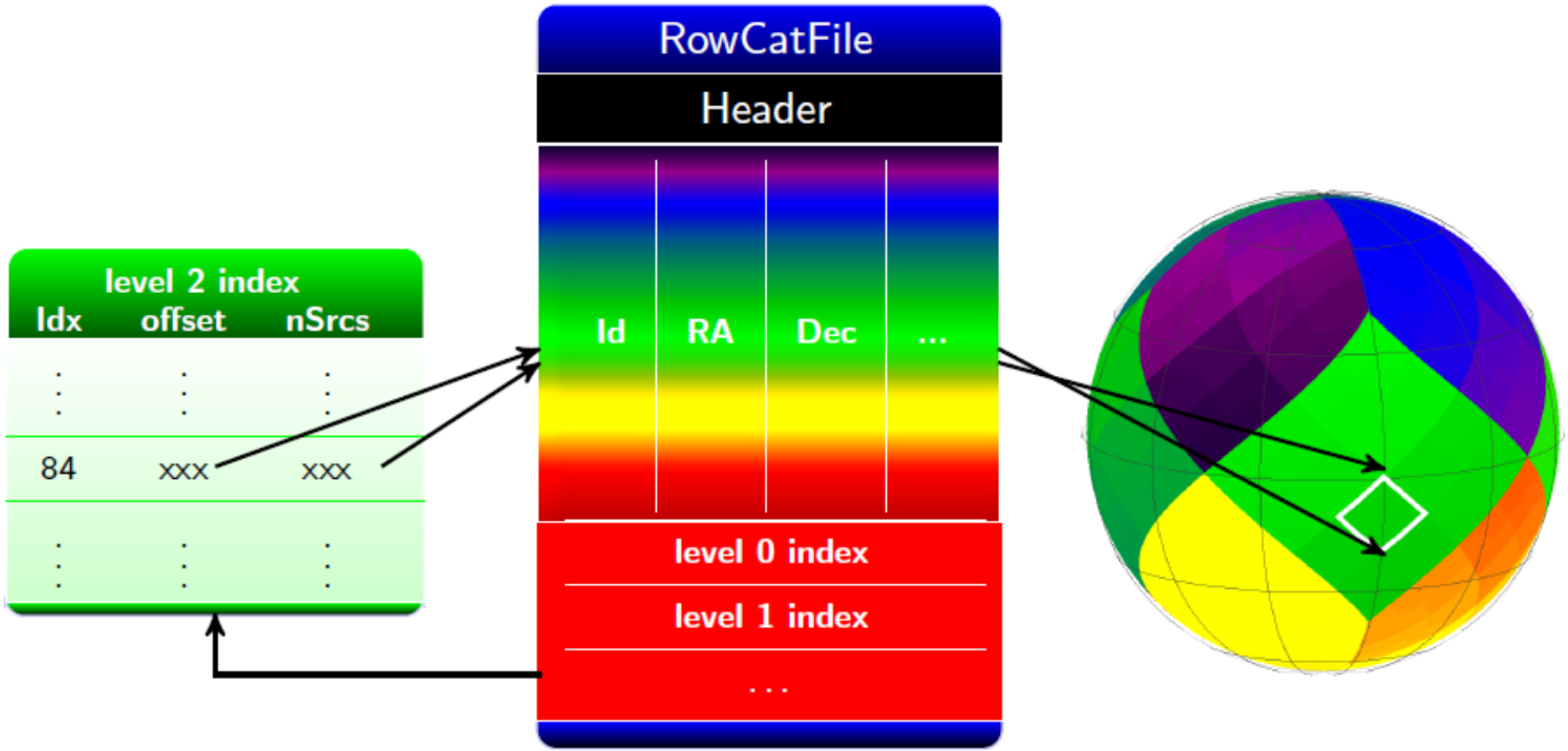
- dedicated scripts like find_allwise.py, find_gaia_dr2.py, etc

```
Query 2mass arround M1: find_2mass.py M1
Query sdss12 arround '217.488910+36.086880': find_sdss-dr12.py "217.488910+36.086880"
```

You can also query with constraints (see –help to list the constraints available (case sensitive))

VizieR: accès scriptable

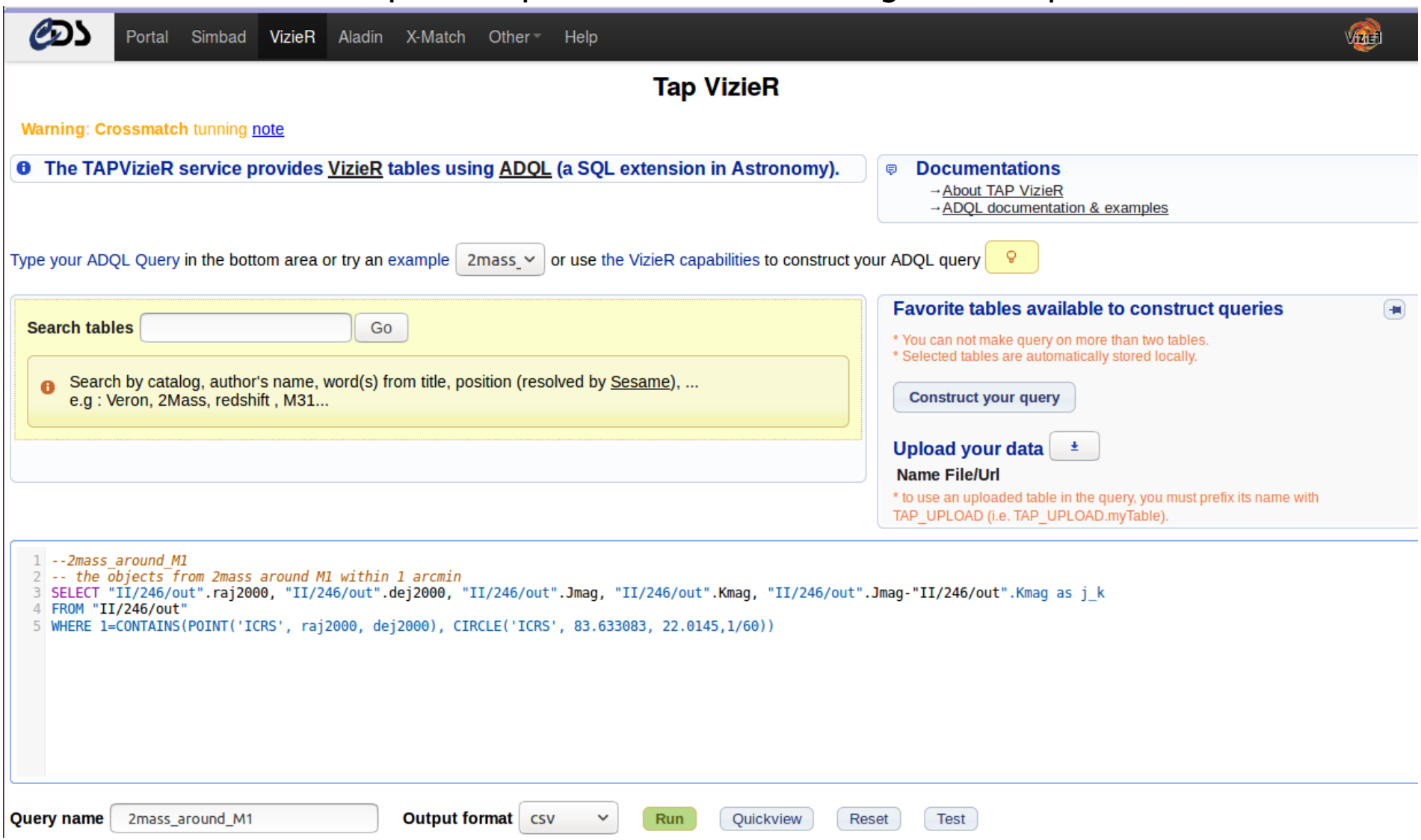
Particularité des grands catalogues: pas en base!

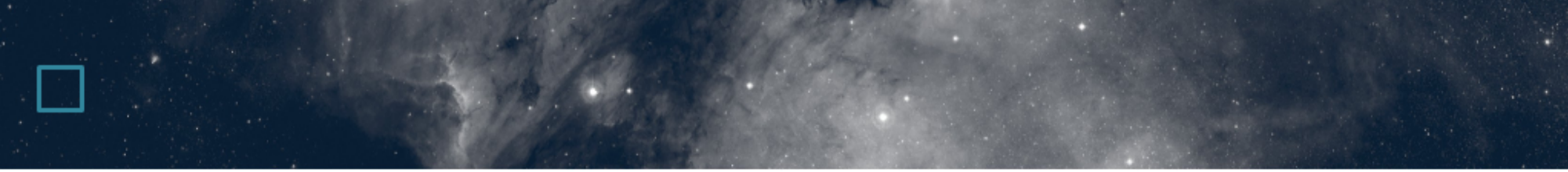


Format binaire spécifique

VizieR: accès scriptable

http://tapvizier.u-strasbg.fr/adql/

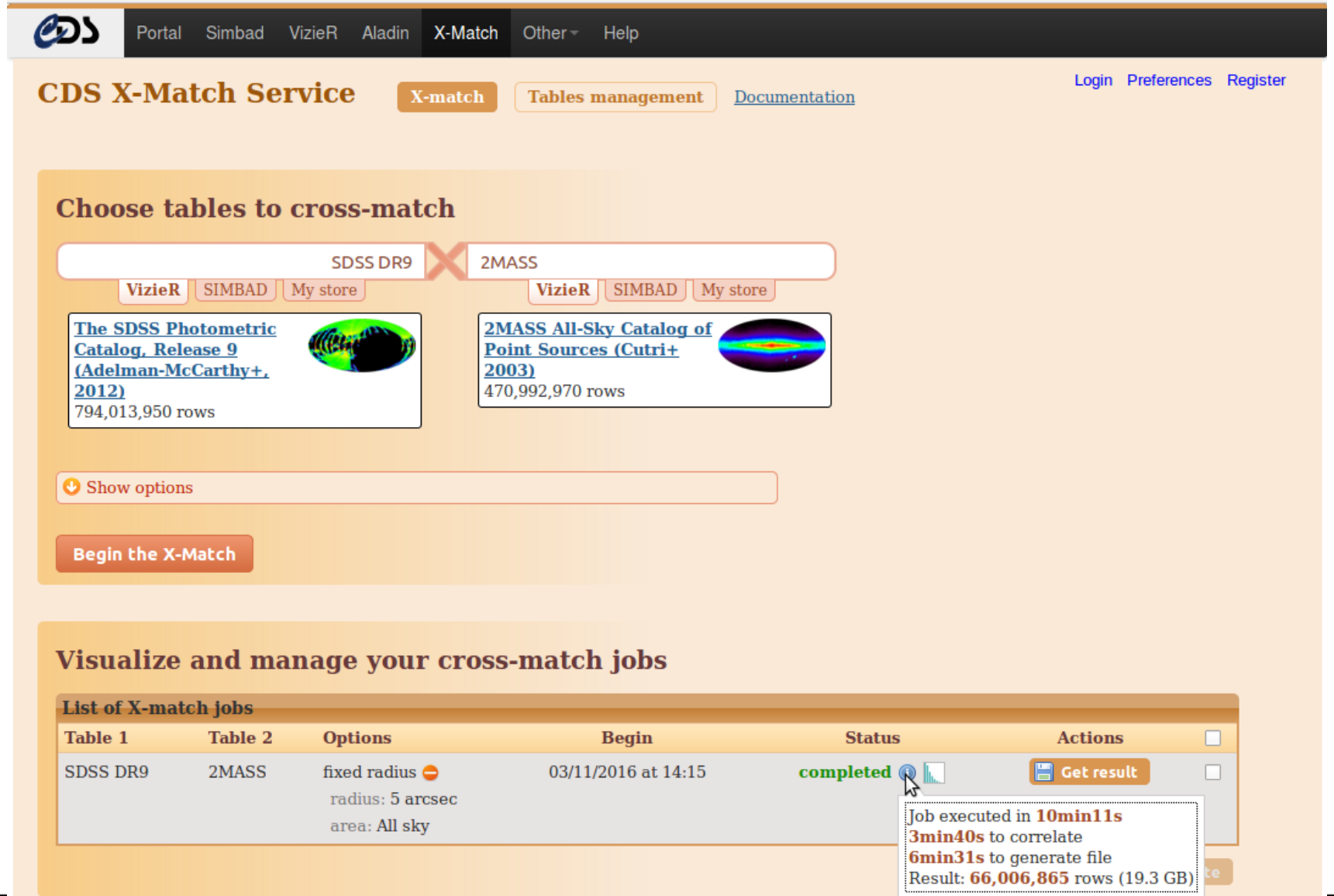




Cross-match simple

XMatch service

http://cdsxmatch.u-strasbg.fr



XMatch service

Example

Using **curl** to match several FITS file with Simbad in Bash

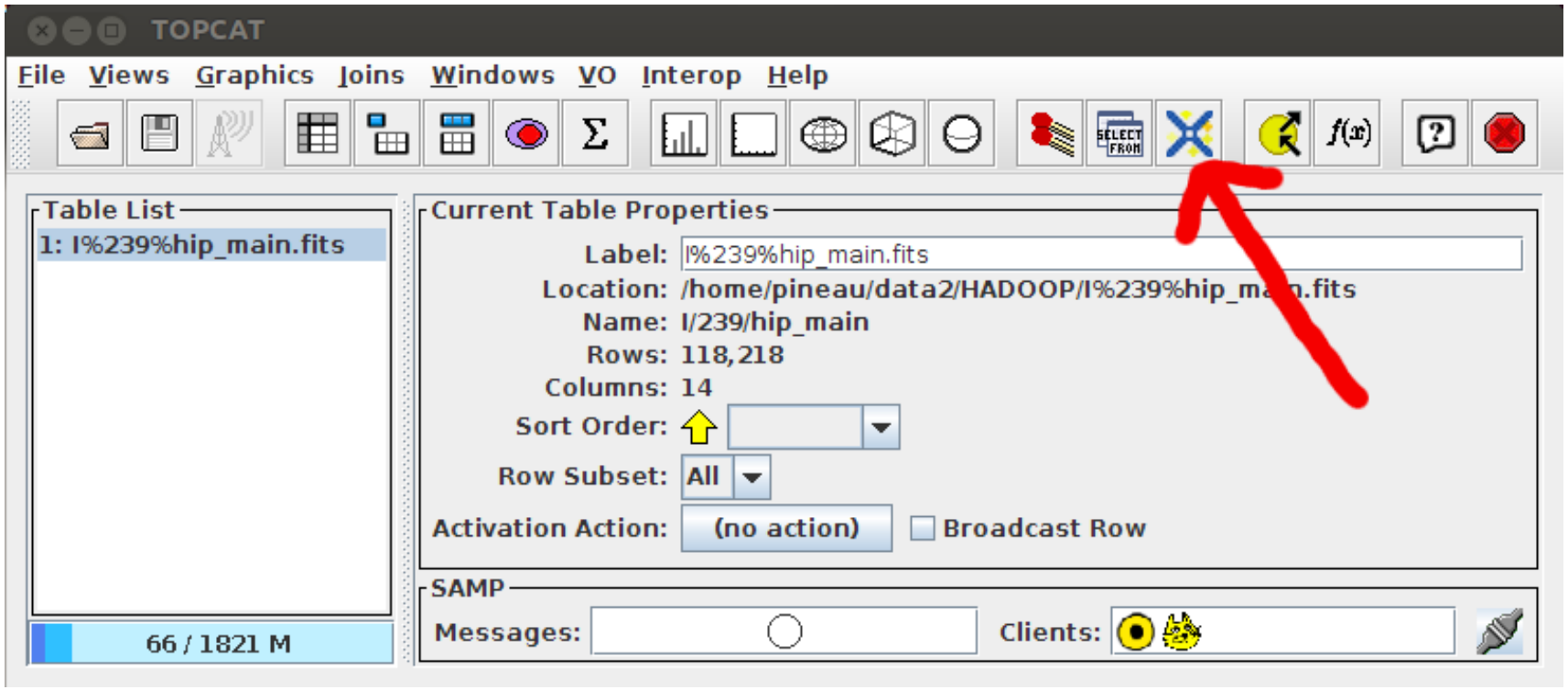
```
for f in file1 file2 file3 file4; do \
    curl -X POST -F request=xmatch \
        -F cat1=@$f.fits -F colRA1=RAJ2000 -F colDec1=DEJ2000 \
        -F cat2=simbad \
        -F distMaxArcsec=25 \
        -F RESPONSEFORMAT=csv \
        http://cdsxmatch.u-strasbg.fr/xmatch/api/v1/sync \
        > $f_vs_simbad_25arcsec.csv \
done
```

Other languages

For Python, Ruby and Java, see here:
`http://cdsxmatch.u-strasbg.fr/xmatch/doc/xmatch-API-usage-examples.html`

XMatch service

http://cdsxmatch.u-strasbg.fr/xmatch/api/v1/sync



Usage statistics

- Web Interface (removing internal usages)

year	#IPs	#Jobs		#Links		Outputs size	
			/day	Billion	M/day	TB	GB/day
2017	1706	8873	**33**	47.9	**179.4**	13.61	**52.2**
2016	1923	11102	30	37.9	104.0	10.1	28.4
2015	1194	7406	20	20.3	55.7	5.0	14.0
2014	1136	5909	16	25.6	70.2	6.6	18.5
2013	1081	5407	14	5.0	13.7	1.2	3.4
2012	535	3699	10	11.5	31.4	2.7	7.5
2011	96	409	7	3.7	67.3	0.83	15.5

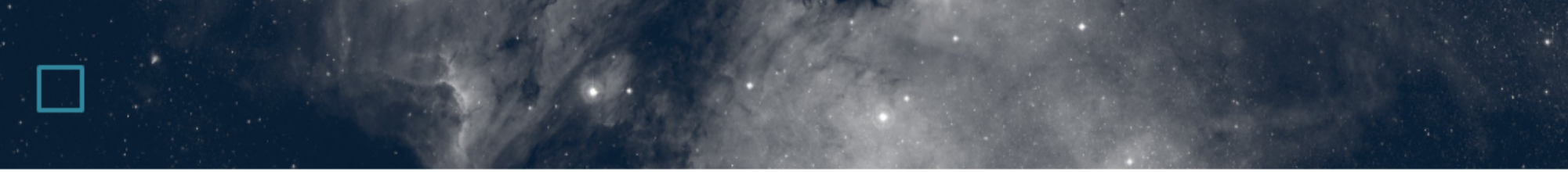
- XXX: computed on incomplete years

Usage statistics

- Synchronous HTTP API (removing internal usages)

year	#IPs	#Jobs	#Links		#Positions (TOPCAT)	
		/day	Billion	M/day	Billion	M/day
2017	1664	**1250**	3.4	**12.2**	4.1	**14.8**
2016	1765	889	2.5	6.7	4.3	11.9
2015	1099	580	2.4	6.6	3.0	8.3
2014	406	49	0.6	1.6	0.3	
2013	46		0.1			

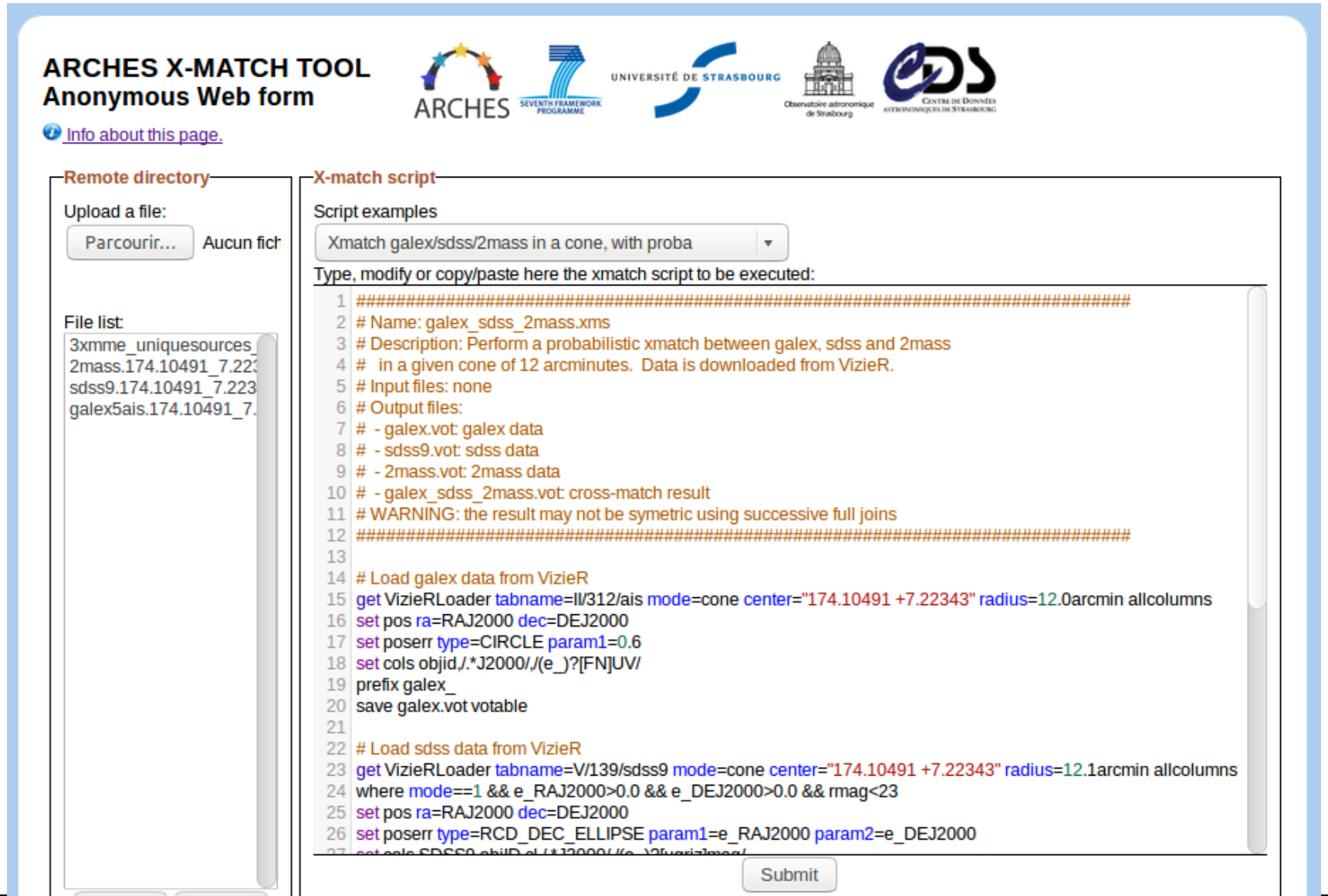
- XXX: computed on incomplete years



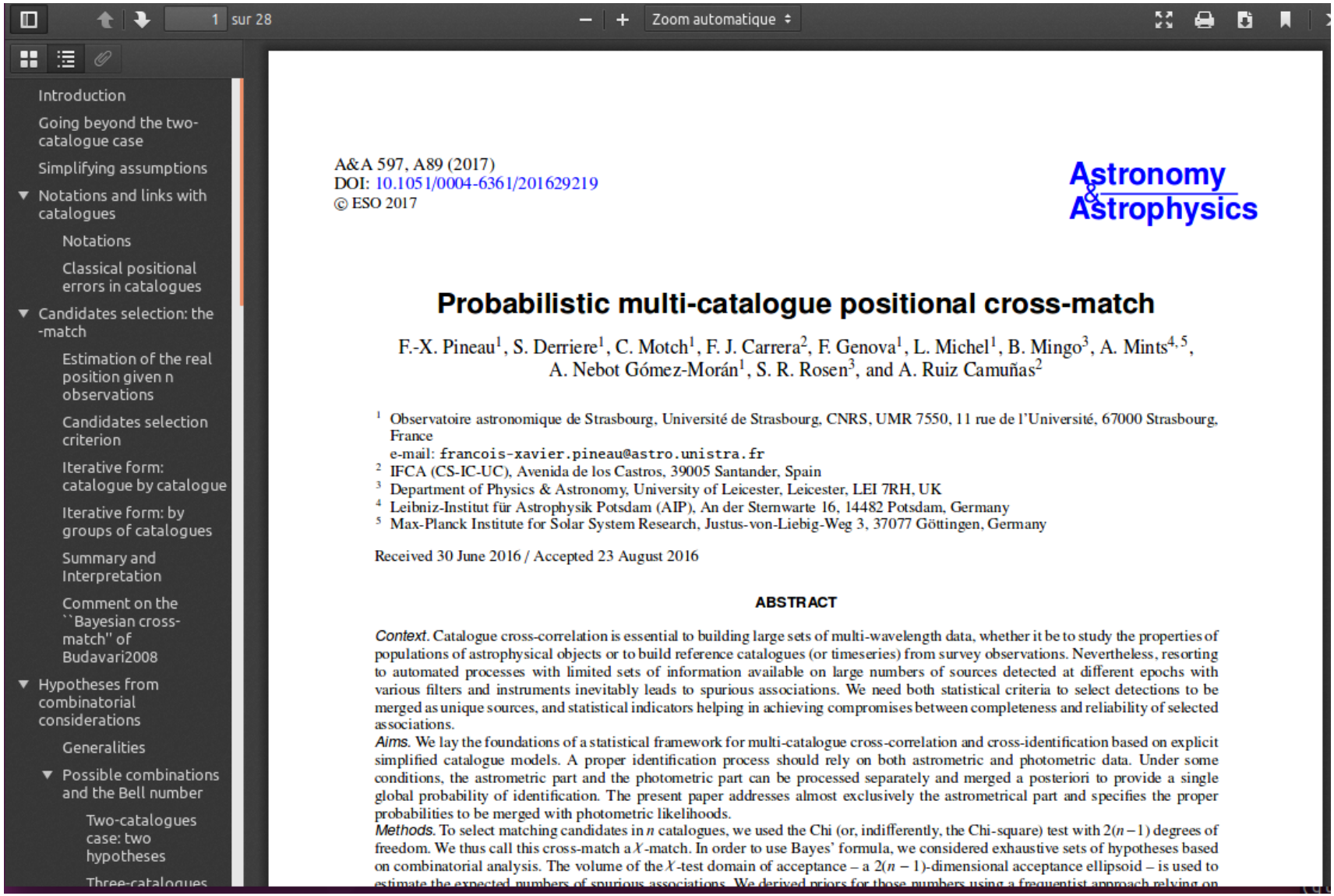
Cross-match complexe

The ARCHES tool

http://serendib.unistra.fr/ARCHESWebService/index.html



The ARCHES tool



1 sur 28 — Zoom automatique

A&A 597, A89 (2017)
DOI: 10.1051/0004-6361/201629219

Astronomy & Astrophysics

Probabilistic multi-catalogue positional cross-match

F.-X. Pineau[1], S. Derriere[1], C. Motch[1], F. J. Carrera[2], F. Genova[1], L. Michel[1], B. Mingo[3], A. Mints[4,5], A. Nebot Gómez-Morán[1], S. R. Rosen[3], and A. Ruiz Camuñas[2]

[1] Observatoire astronomique de Strasbourg, Université de Strasbourg, CNRS, UMR 7550, 11 rue de l'Université, 67000 Strasbourg, France
e-mail: francois-xavier.pineau@astro.unistra.fr
[2] IFCA (CS-IC-UC), Avenida de los Castros, 39005 Santander, Spain
[3] Department of Physics & Astronomy, University of Leicester, Leicester, LEI 7RH, UK
[4] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[5] Max-Planck Institute for Solar System Research, Justus-von-Liebig-Weg 3, 37077 Göttingen, Germany

Received 30 June 2016 / Accepted 23 August 2016

ABSTRACT

Context. Catalogue cross-correlation is essential to building large sets of multi-wavelength data, whether it be to study the properties of populations of astrophysical objects or to build reference catalogues (or timeseries) from survey observations. Nevertheless, resorting to automated processes with limited sets of information available on large numbers of sources detected at different epochs with various filters and instruments inevitably leads to spurious associations. We need both statistical criteria to select detections to be merged as unique sources, and statistical indicators helping in achieving compromises between completeness and reliability of selected associations.

Aims. We lay the foundations of a statistical framework for multi-catalogue cross-correlation and cross-identification based on explicit simplified catalogue models. A proper identification process should rely on both astrometric and photometric data. Under some conditions, the astrometric part and the photometric part can be processed separately and merged a posteriori to provide a single global probability of identification. The present paper addresses almost exclusively the astrometrical part and specifies the proper probabilities to be merged with photometric likelihoods.

Methods. To select matching candidates in n catalogues, we used the Chi (or, indifferently, the Chi-square) test with $2(n-1)$ degrees of freedom. We thus call this cross-match a χ-match. In order to use Bayes' formula, we considered exhaustive sets of hypotheses based on combinatorial analysis. The volume of the χ-test domain of acceptance – a $2(n-1)$-dimensional acceptance ellipsoid – is used to estimate the expected numbers of spurious associations. We derived priors for those numbers using a frequentist approach relying on

The ARCHES tool

Algorithm	param	#tbl	prop.mot.	index struct.
chi2 (χ^2)	proba	2	l[1], r[2], b[3]	M/TM-tree
proba2_v*x*	proba	2	no (?)	M-tree
proba3_v*x*	proba	3	no (?)	M-tree
proba4_v*x*	proba	4	no (?)	M-tree
probaN_v*x*	proba	n	no (?)	M-tree
knn	k+dist	2	r, b	kd/M/TM-tree
cone	dist	2	l, r, b	kd/M/TM-tree
mec[1]	dist	n	no (?)	kd/M-tree
ext_l[1]	r	2	no	M-tree
ext_r[2]	r	2	no	M-tree
ext_b[3]	r	2	no	M-tree
...	...	...	...	...

XMatches are chainable: 1 χ^2 xmatch of 4 tables = 3 χ^2 xmatches of 2 tables!
4 to 11 joins ($LIRF\bar{L}\bar{I}\bar{R}L'I'R'F'$) are supported according to the algorithm.

- [1] l: left table contains extended objects or proper motions;
- [2] r: right table contains extended objects or proper motions;
- [3] b: both left and right tables contain extended objects or proper motion

Hypotheses

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

Hypotheses

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

Hypotheses

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis

Hypotheses

To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - A_BC
    - A_B_C

Hypotheses

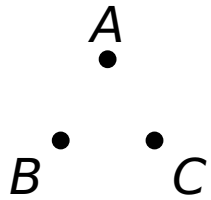
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



Hypotheses

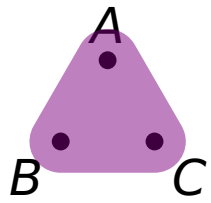
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



Hypotheses

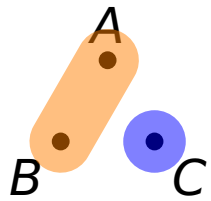
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



Hypotheses

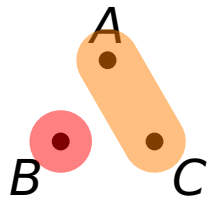
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



Hypotheses

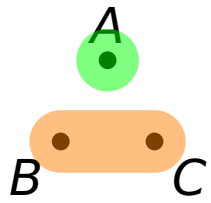
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



Hypotheses

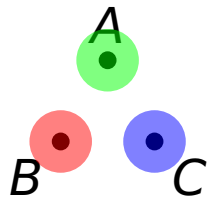
To compute Bayes probabilities, we MUST consider all possible hypothesis.

- Law of total probabilities:

$$\sum_{i=1}^{n} p(H_i) = 1$$

- For 2 catalogues
  - 2 hypothesis
    - AB (H_0)
    - A_B

- For 3 catalogues
  - 5 hypothesis
    - ABC (H_0)
    - AB_C
    - AC_B
    - A_BC
    - A_B_C



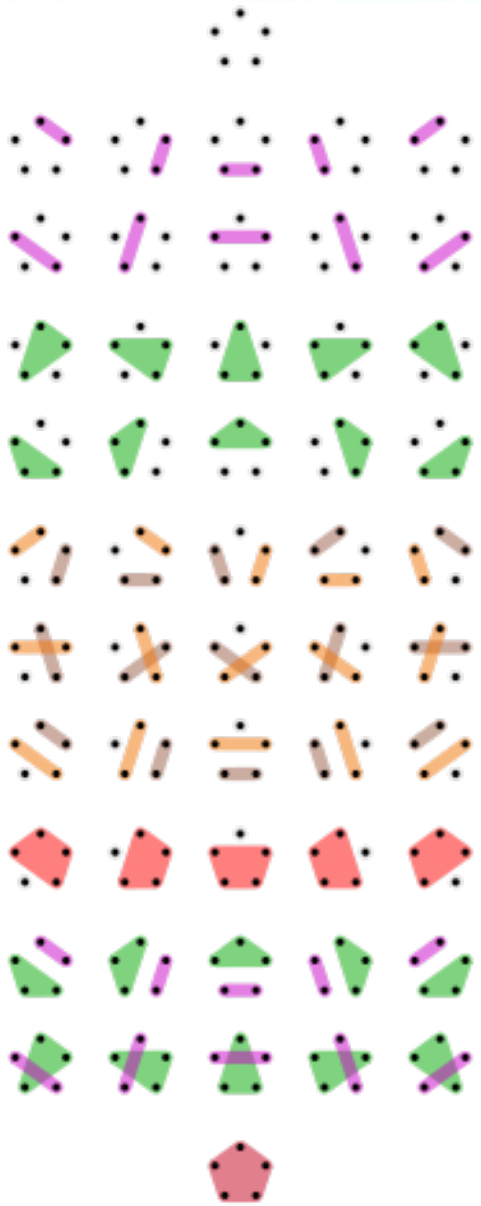
Hypotheses

- We generalised for n catalogues
- The number of hypothesis to be tested is given by the BELL number

Table : Values of the seven first BELL numbers

n	2	3	4	5	6	7
B_n	2	5	15	52	203	877

  - n number of catalogues
  - n=5 catalogues $\rightsquigarrow$ 52 probabilities to be computed
- $\Rightarrow$ Combinatorial explosion!

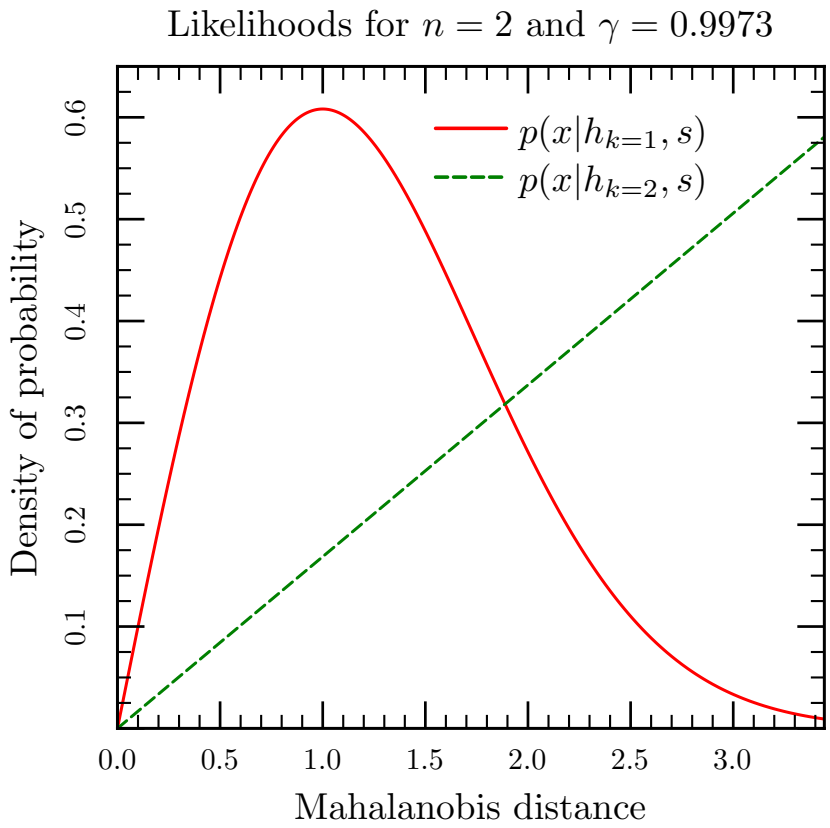


Credits: wikipedia

Bayesian probabilities

Summary for 2 catalogues

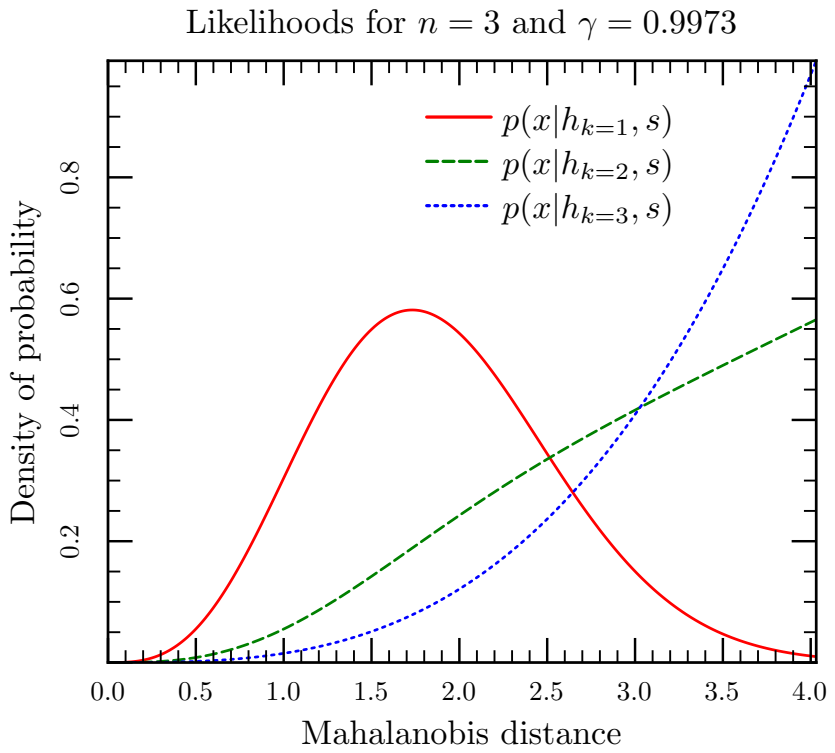
- 2 hypotheses
- 2 likelihoods
  - $H_0 = AB$: $p(x|H_{AB})$, Chi of 2 dof
  - $H_1 = A\_B$: $p(x|H_{A\_B})$, Poisson 2D
- 2 priors based on geometrical estimates



Bayesian probabilities

For 3 catalogues

- 5 hypotheses
  - *ABC*, 1 real source
  - *AB_C*, 2 real sources
  - *AC_B*, 2 real sources
  - *A_BC*, 2 real sources
  - *A_B_C*, 3 real sources
- 3 likelihoods
  - 1 likelihood by number of real source
  - $p(x|H_{ABC}) = \chi_{dof=4}(x)/\gamma$: Chi 4 dof
  - $p(x|H_{AB\_C}) = p(x|H_{AB\_C}) = p(x|H_{AB\_C})$
  - $p(x|H_{A\_B\_C}) = 4x^3/k_\gamma^4$: Poisson 4D



Prise en compte de paramètres photométriques

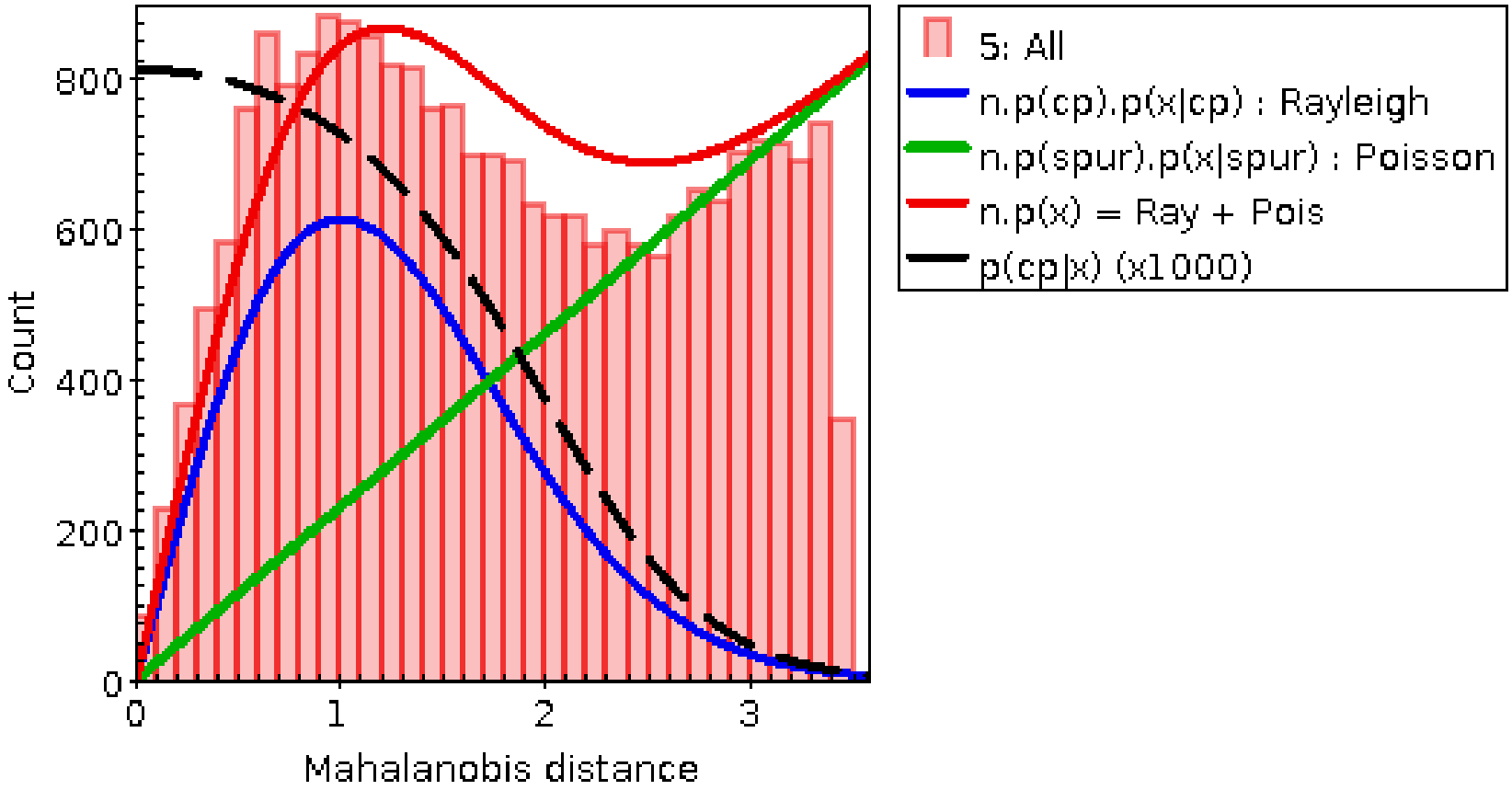
XMatch example

- Roughly reproducing Salvato et al. (2018) results (J/MNRAS/473/4937/xmmslew2) with the ARCHES tool + CDS classification (prototype) service
- Write and ARCHES cross-match script

```
# Load and set the XMM data to be cross-matched
get FileLoader file=XMMSL2_exgal_fewcol_2017APR12.fits
where RADEC_ERR < 10.0
set pos ra=RA dec=DEC
set poserr type=CIRCLE param1=RADEC_ERR/sqrt(2)
set cols *
prefix x
# Load and set the AllWISE data to be cross-match
get FileLoader file=candidate_ALLWISE_counterparts_unique_2017APR12.fits.gz
where eeMaj < 0.75
set pos ra=RA dec=DEC
set poserr type=ELLIPSE param1=eeMaj param2=eeMin param3=eePA
set cols *
prefix w
# Perform the cross-match, add the angular distance and save the result
xmatch probaN_v1 joins=I completeness=0.9973 area_w=0.01851769294883401575
               area_x=0.01851769294883401575 area_xw=0.01388
merge dist mec
save xmmslew2_vs_allwise.fits fits
```

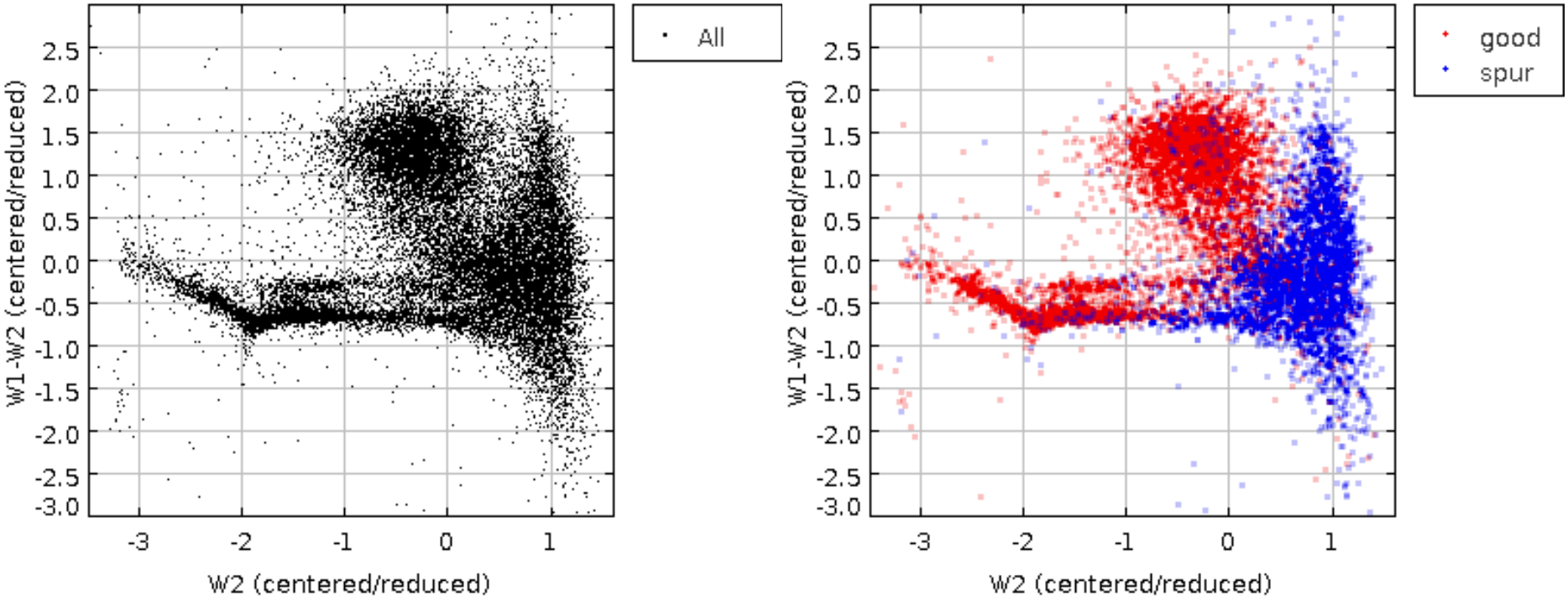
NWAY / ARCHES: result

- Cross-match result: 23 813 associations
- Number of spurious matches ($n.p(spur)$) overestimated (?)



NWAY / ARCHES: photometry

- From Salvato et al. (2018): W2 vs W1-W2
- Learning samples arbitrary defined:
  - Good: d<6" && RADEC_ERR<8 && proba_xw>0.75
  - Spurious: d>13" && proba_xw<0.05



Classification service

- 3 CSV files: all matches, good matches, spurious matches
- Each file contains: $id, w2, w1 - w2$
- Using the CDS prototype service:

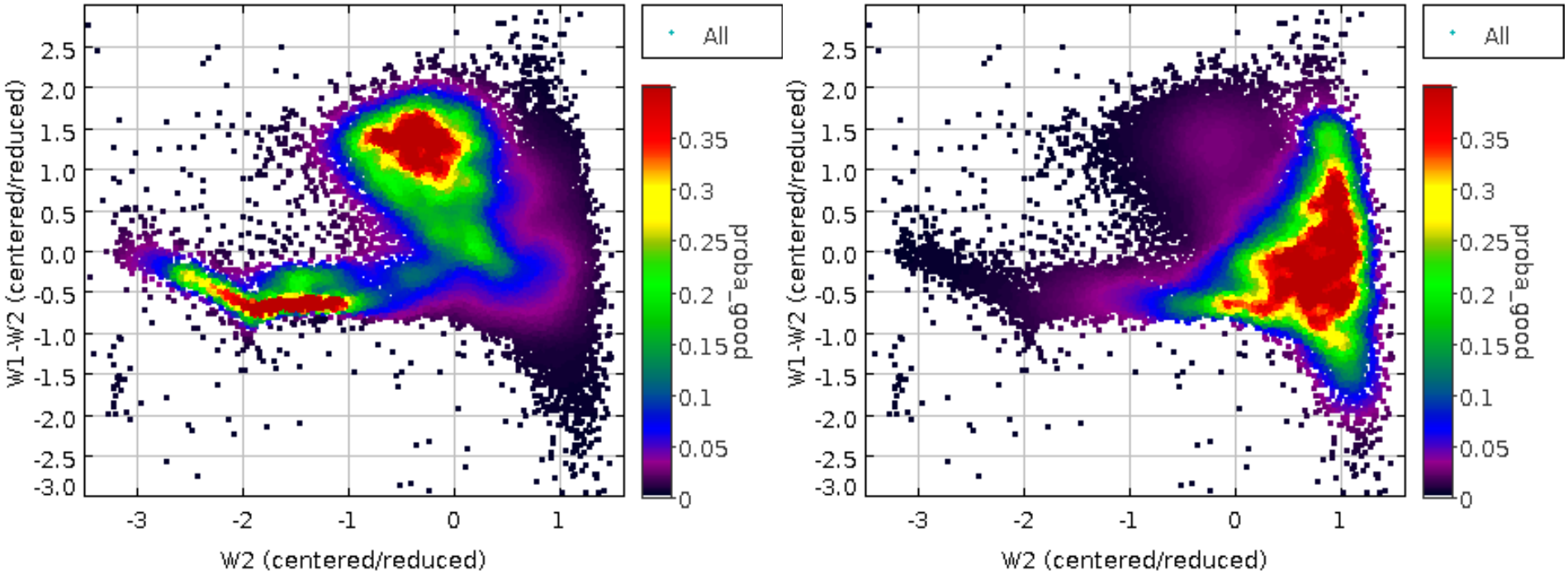
```
# Put the data files into the distant server
./classif.bash put good slew_vs_allwise.good.csv # 4565 rows
./classif.bash put spur slew_vs_allwise.spur.csv # 3082 rows
./classif.bash put data slew_vs_allwise.all.csv  # 23799 rows
# Performs the classification of the data and save the result
./classif.bash kdc samplepoint -k 75 -p good:0.425\;spur:0.575 -ho > result.csv
# Ask for the confusion matrix by self-classifying the LS
./classif.bash kdc samplepoint -k 75 -p good:0.425\;spur:0.575 -cr
```

- Confusion matrix:

actual \predicted	good	spurious
good	85.96%	14.04%
spurious	9.73%	90.27%

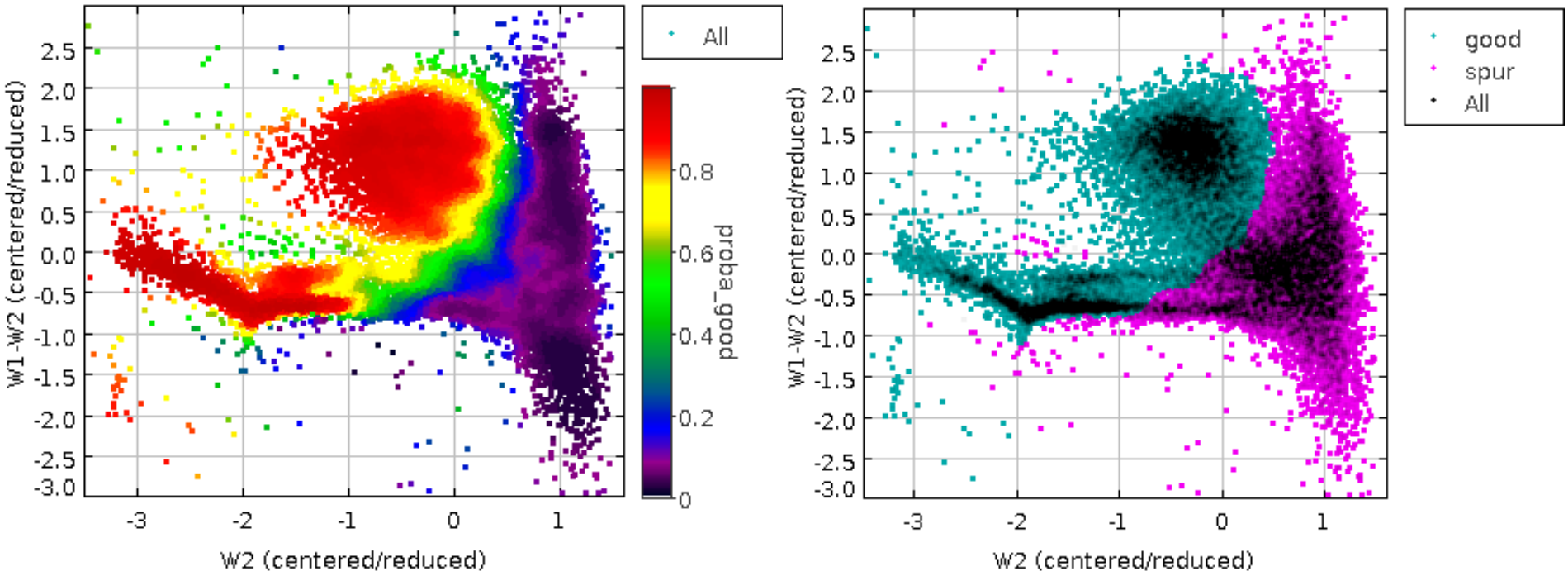
NWAY / ARCHES: photometry

- Likelihoods (distributions) computed by kernel smoothing (sample point estimator, k=75)
  - left: $p(\vec{m}|good)$
  - right: $p(\vec{m}|spur)$



NWAY / ARCHES: photometry

- Left: classification result $p(good)$
- Right: binary classification Good/Spurious ($p(good) > 0.5$, $p(good) < 0.5$)



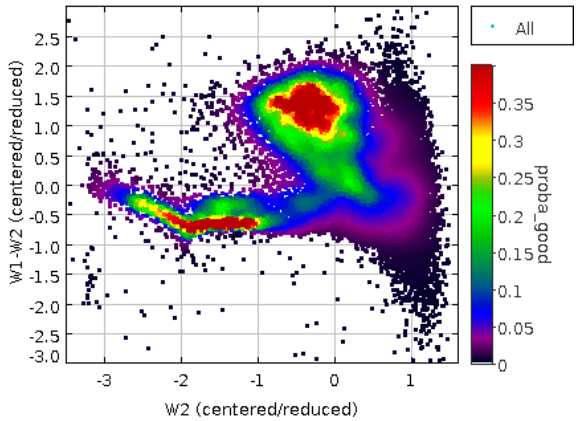
Merging with positional proba

- Accounting for photometric likelihoods (simplified Eq. 154 of Pineau et al. 2017)

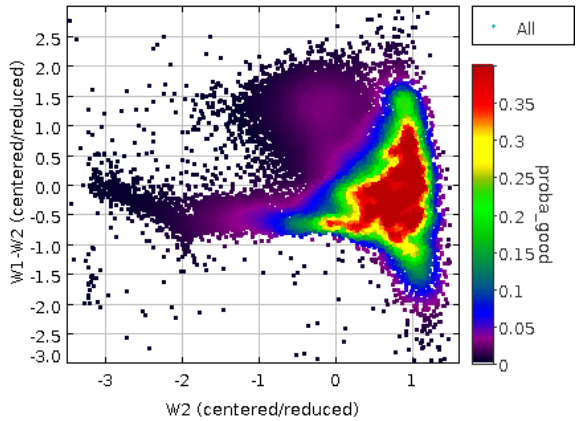
$$p(real|x,\vec{m}) = \frac{p(real|x)p(\vec{m}|real)}{p(real|x)p(\vec{m}|real) + (1-p(real|x))p(\vec{m}|spur)}$$

$p(real|x)$ = purely positional probability

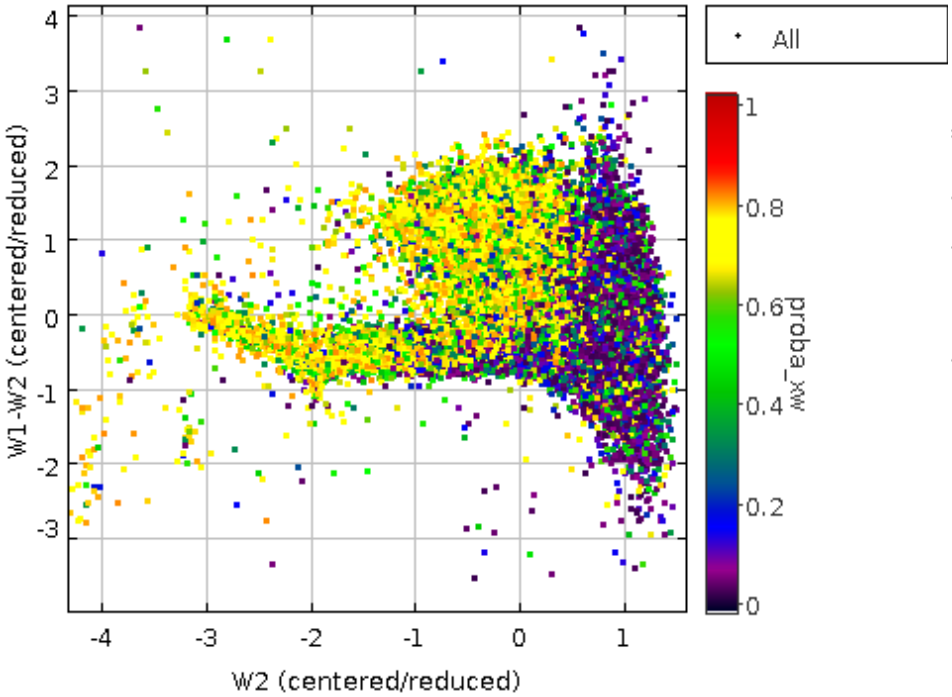
$p(\vec{m}|real) =$

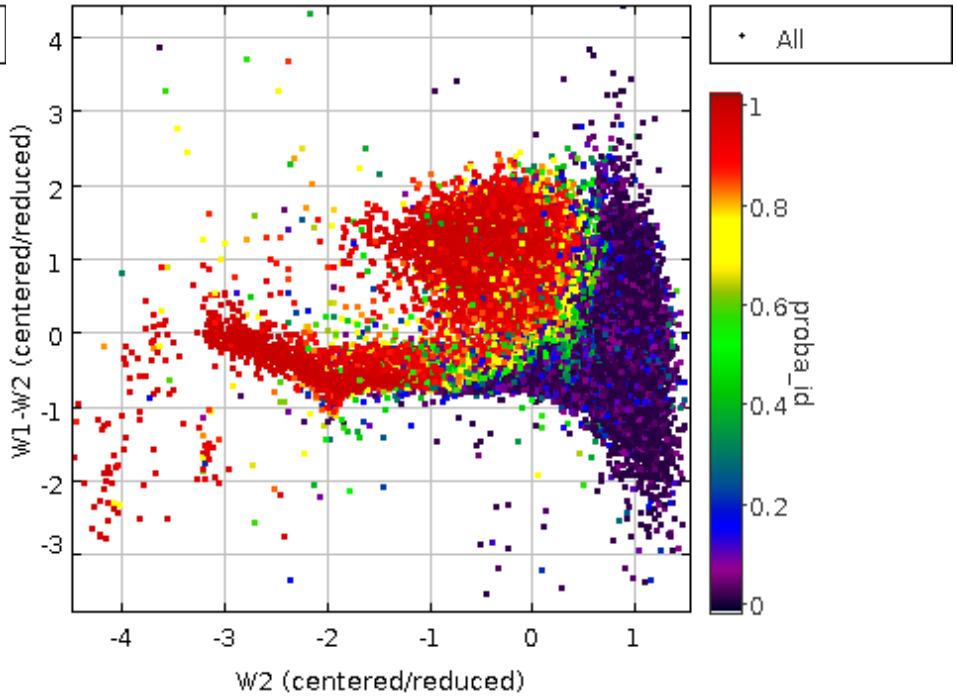


$p(\vec{m}|spur) =$



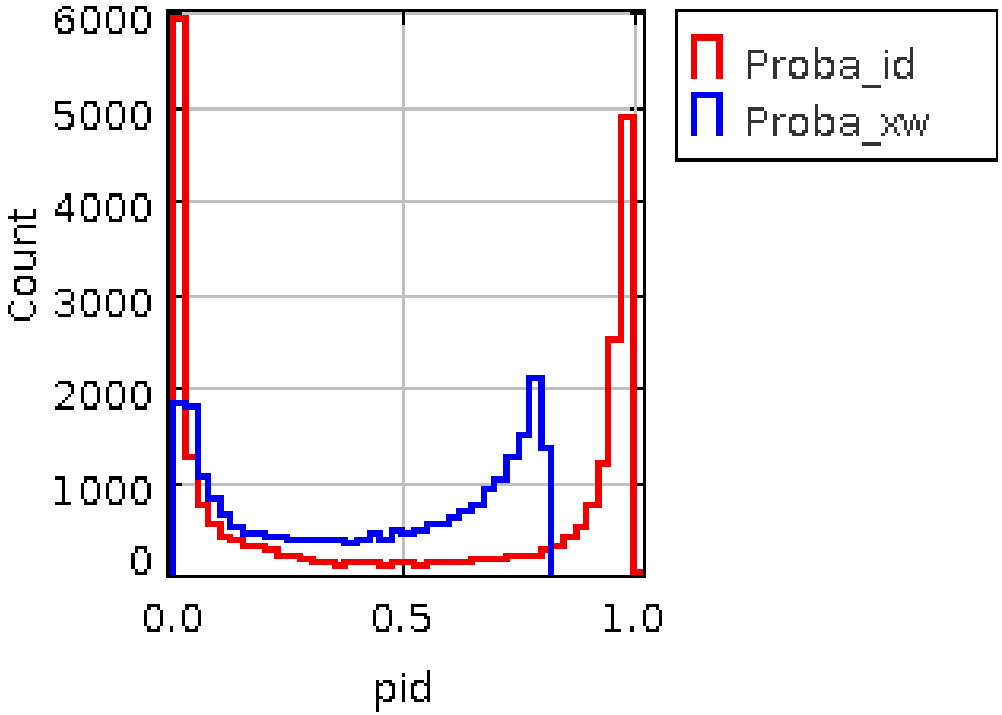
NWAY / ARCHES: photometry

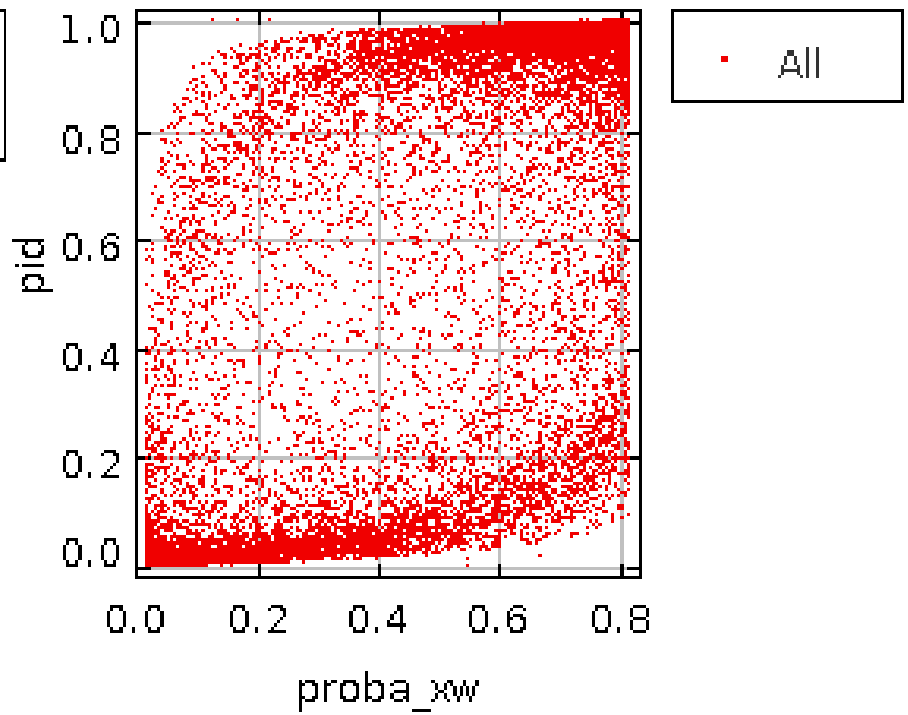




- Left: positional probabilities
- Right: probabilities accounting for photometric likelihoods

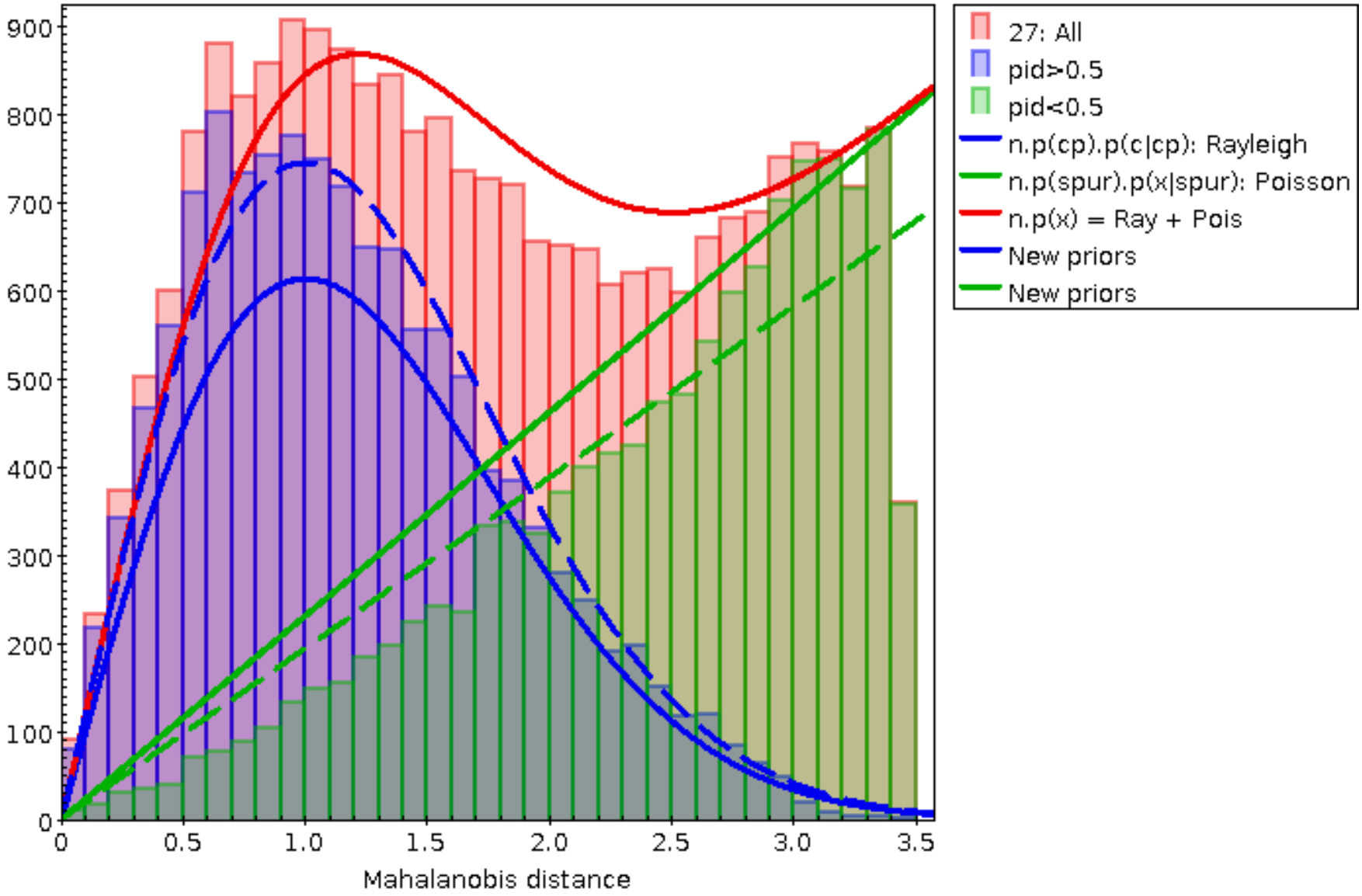
NWAY / ARCHES: proba id





- Clearer separation of low and high probabilities

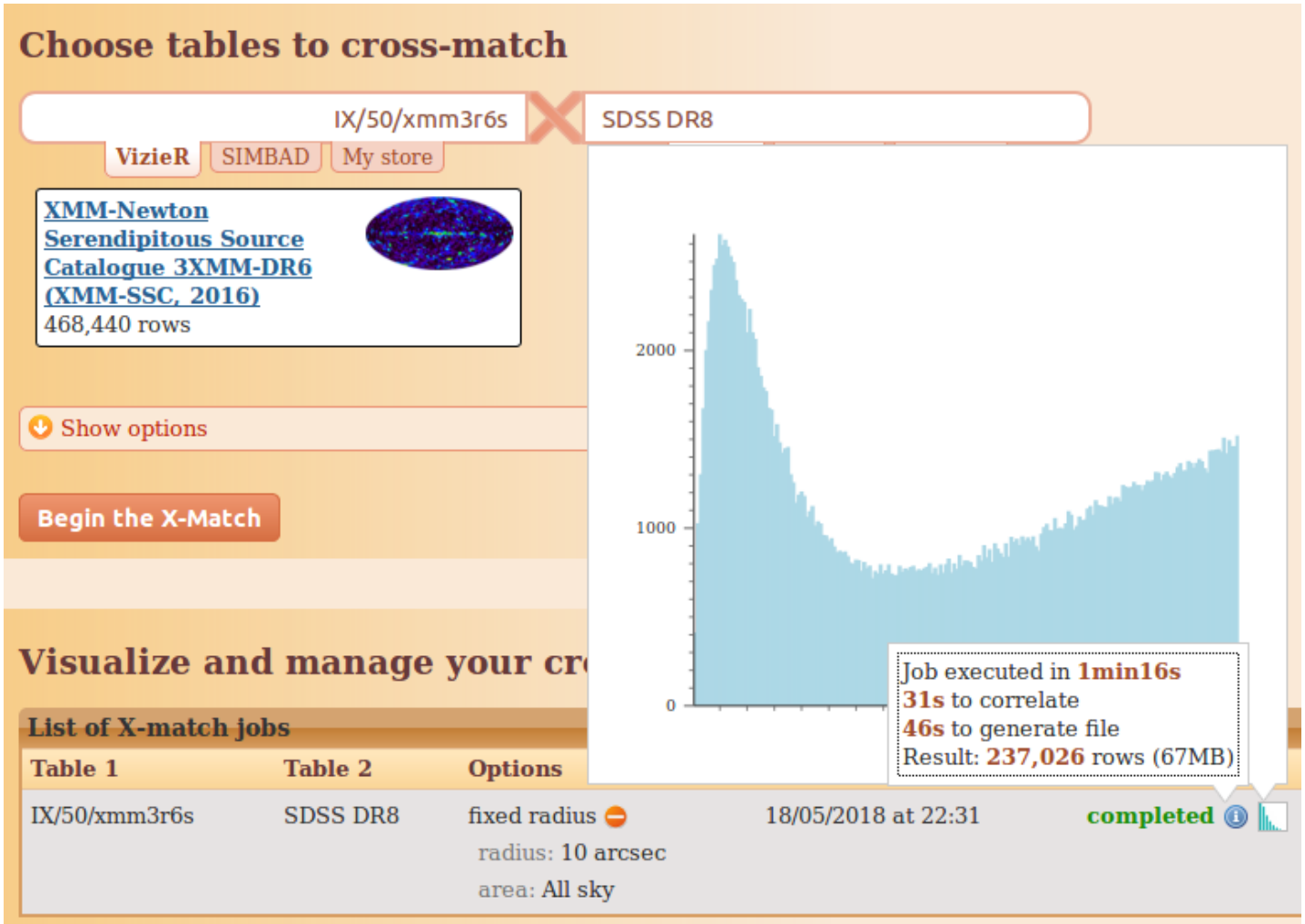
NWAY / ARCHES: proba id



XMM vs SDSS DR8

Testing the same method to cross-match XMM and SDSS

XMM vs SDSS DR8



http://cdsxmatch.u-strasbg.fr

XMM vs SDSS DR8

- Quick-and-dirty test in 4 dimensions
  - Simple 10 arcsec cross-match
  - Keep only *primary unresolved* objects having a *clean photometry*
  - Mahalanobis distance: $d_\sigma \approx \frac{d}{\sqrt{\left(\frac{SC\_POSERR}{\sqrt{2}}\right)^2 + RA\_ERR \times DE\_ERR}}$
  - Lazy learning samples definition:
    - ★ 19 676 “real” associations: $d < 1''$ && $d_\sigma < 1.5$
    - ★ 7 784 “spurious” associations: $d > 8''$ && $d_\sigma > 6$
  - User defined prior $p(cp)$ going to $d_{\sigma,max} = 5$
  - From Eq. 149 of Pineau et al. (2017):

$$p(cp|d_\sigma) = \frac{1}{1 + \frac{1-p(cp)}{p(cp)} \frac{2}{d_{\sigma,max}^2 e^{-\frac{d_\sigma^2}{2}}}}$$

XMM vs SDSS DR8

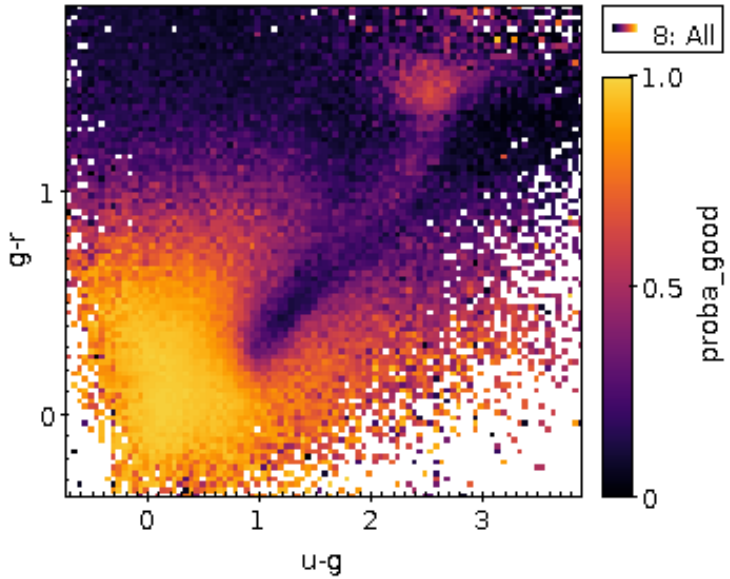
- Using the CDS prototype service:

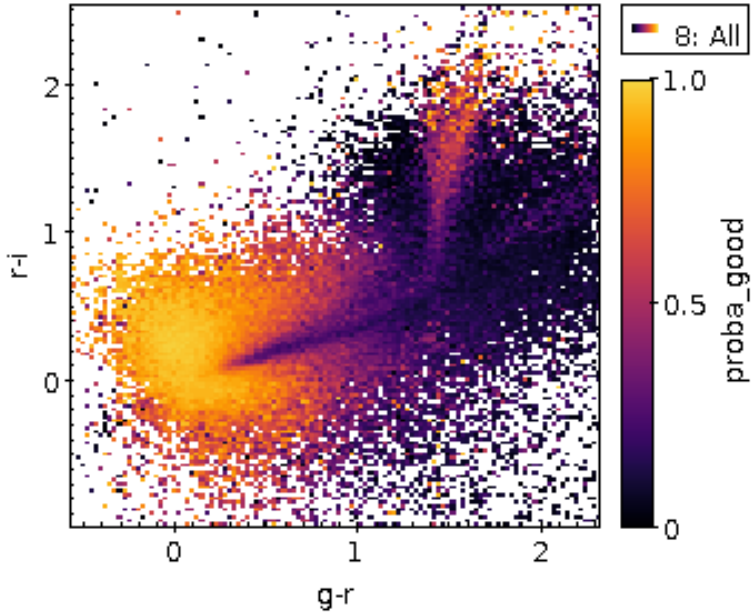
```
# Put the data files into the distant server
./classif.bash put good xmm_sdss8.unres.good.csv # 19676 rows
./classif.bash put spur xmm_sdss8.unres.spur.csv # 7784 rows
./classif.bash put data xmm_sdss8.unres.all.csv
# Performs the classification of the data and save the result
./classif.bash kdc samplepoint -k 25 -p good:0.55\;spur:0.45 -ho > result.csv
# Ask for the confusion matrix by self-classifying the LS
./classif.bash kdc samplepoint -k 25 -p good:0.55\;spur:0.45 -cr
```

- Confusion matrix:

actual \predicted	good	spurious
good	86.91%	13.09%
spurious	12.28%	87.72%

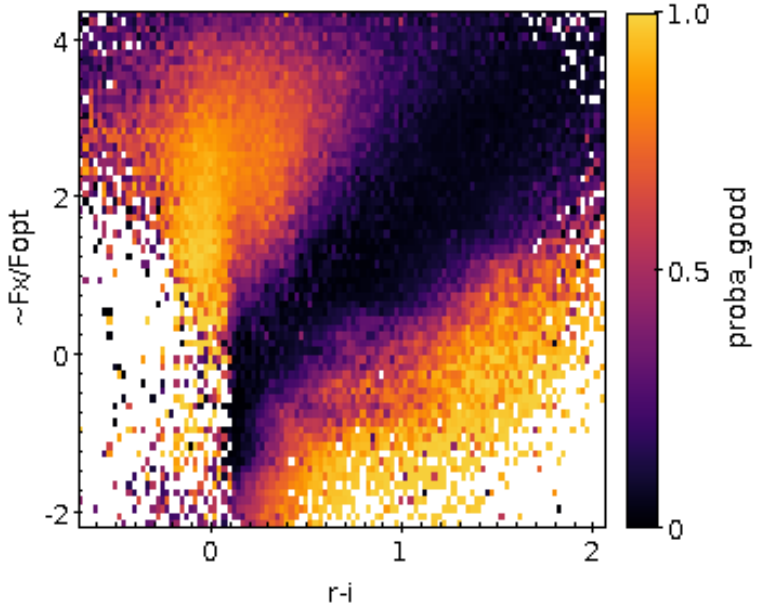
XMM vs SDSS DR8





Mean of the 4D KDC output probabilities in

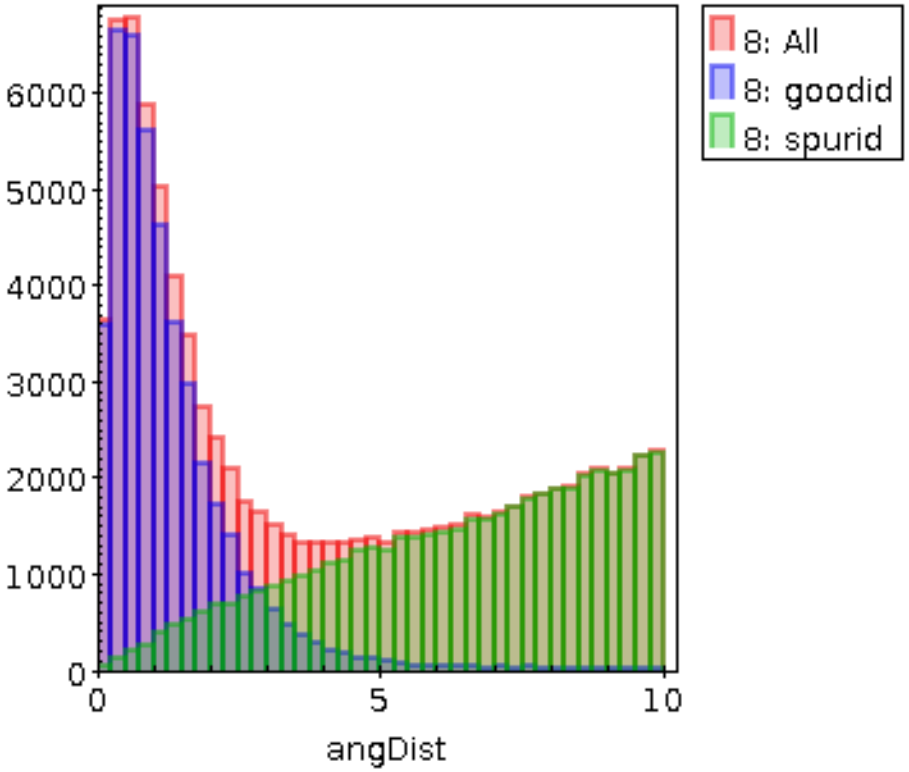
- $u - g$ vs $g - r$
- $g - r$ vs $r - i$
- $r - i$ vs $\propto F_X/F_r$

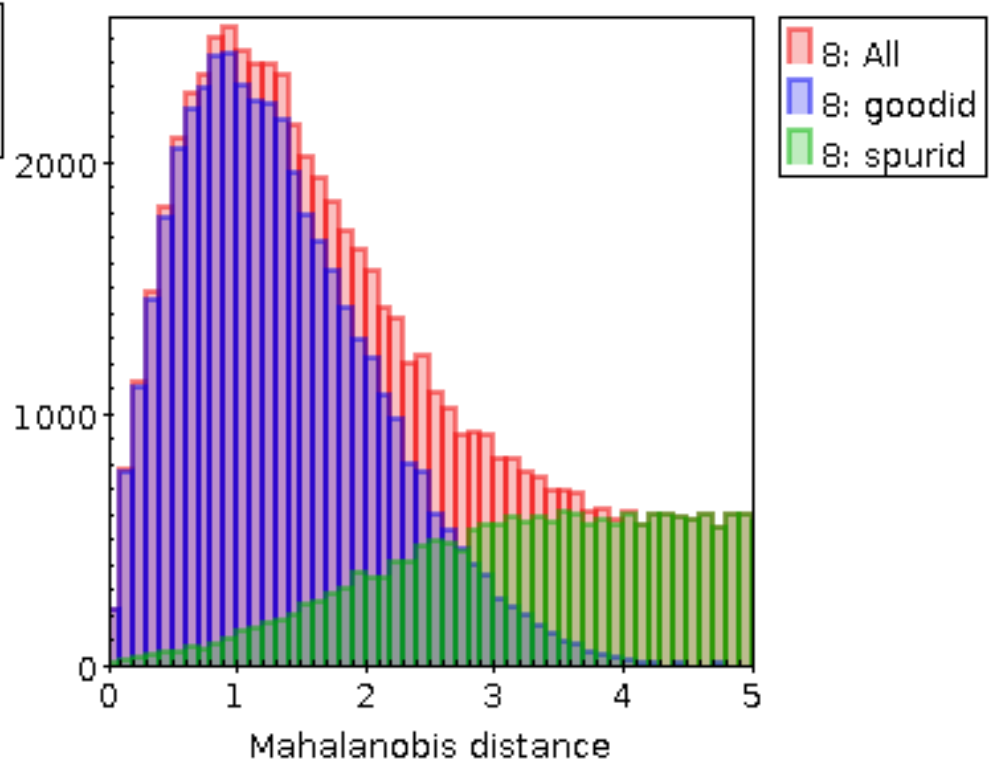


XMM DR7 vs SDSS DR8

Using 4D photometric likelihoods to compute final proba $p(cp|d_\sigma, \vec{m})$, and considering:

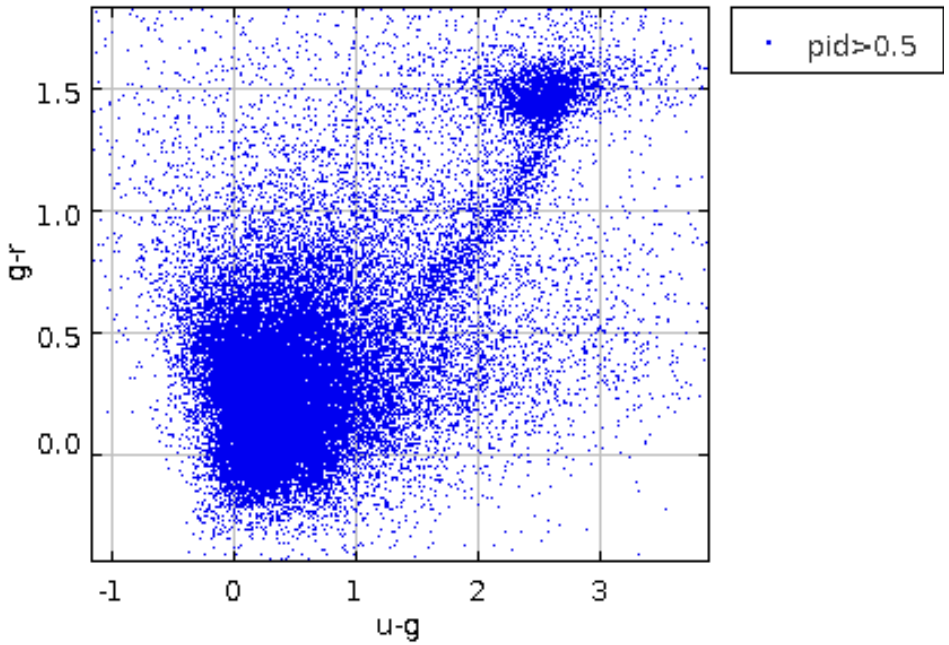
- real matches as $p(cp|d_\sigma, \vec{m}) > 0.5$
- spurious matches as $p(cp|d_\sigma, \vec{m}) < 0.5$

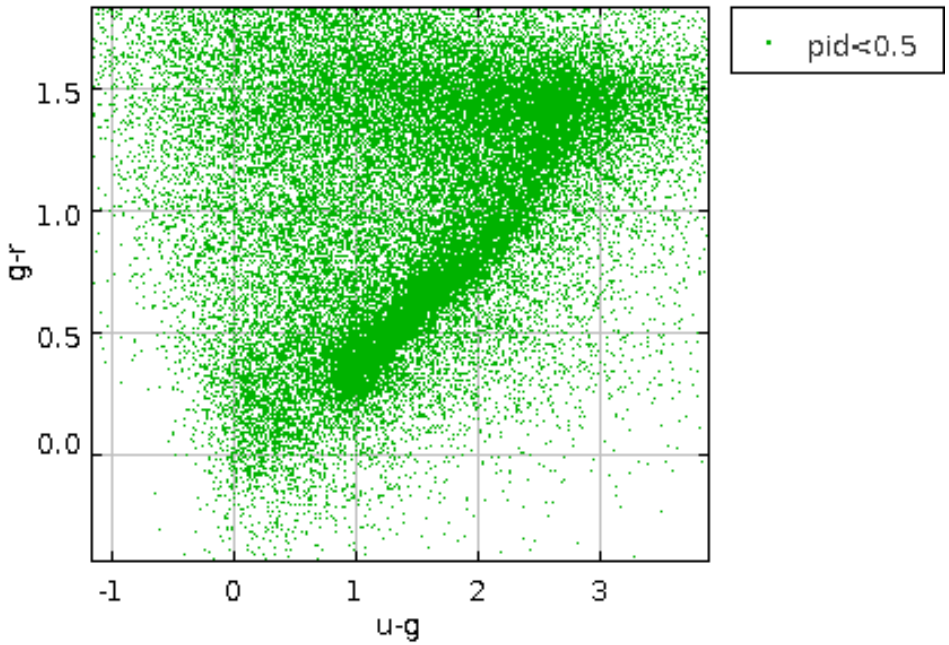


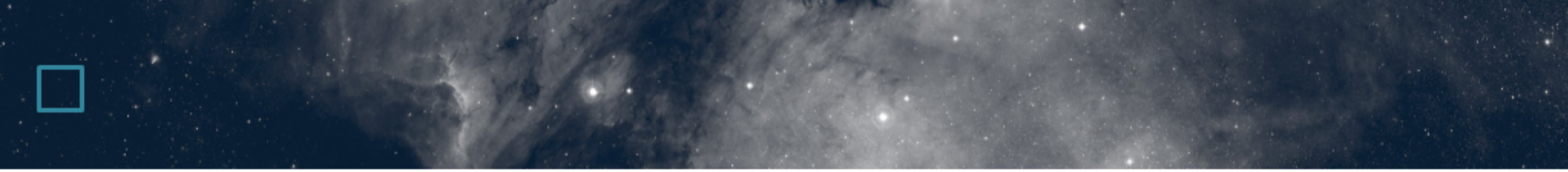


XMM DR7 vs SDSS DR8

- $u - g$ vs $g - r$ diagrams of estimated as real and estimated as spurious associations.







Travail exploratoire

Index spatio-temporel

- Coordonées sphériques + axe temporel entremêlés
  - HEALPix + axe auxiliare (Z-order curve 3D)
- But:
  - répondre rapidement à des requêtes spatio-temporelle
  - décrire des regions spatio-temporelles avec des MOC
- Implémenté: $(\alpha, \delta, t) \rightarrow idx$, $idx \leftarrow center(\alpha, \delta, t)$
- A implémenter: cone search + time range $\rightarrow$ MOC
- A tester: taille des regions (MOC) en pratique

Perspective

Vers le Big Data (Spark)?
Test preliminaire:

- Cross-match Gaia DR1 (1.1 G srcs) vs IGSL3 (1.3 G srcs)
- 5 arcsec $\rightsquigarrow$ 1.6 G associations
- Cluster de 9 machines reformées $\rightsquigarrow$ 10 min
- Algorithme amélioré (kd-tree) à tester