

# Cross-identification au CDS

François-Xavier Pineau<sup>1</sup>, Thomas Boch<sup>1</sup>, Sébastien Derrière<sup>1</sup>, Gilles Landais<sup>1</sup>,  
André Schaaff<sup>1</sup>, Noémie Wali<sup>2</sup>

<sup>1</sup>Observatoire Astronomique de Strasbourg, Université de Strasbourg, CNRS

<sup>2</sup>Université de Technologie de Belfort-Montbéliard (UTBM)

Réunion annuelle de l'Action Spécifique Observatoires Virtuels France  
Institut d'Astrophysique de Paris, le 15 mars 2016



# Sommaire

- 1 Simple x-match
- 2 Multi-catalogue  $\chi$ -match & x-id

# X-Match within the VO

## Using VO tools

- No dedicated VO protocol
- Topcat
  - ▶ Use the VO Registry and download a VOTable
  - ▶ Perform one Cone Search by row
    - ★ good for VizieR's stats...
    - ★ ...but saturates the server when multi-threaded!
- Topcat / Aladin
  - ▶ Use the VO Registry and download VOTables
  - ▶ Perform the x-match locally
    - ★ Tycho2 vs 2MASS
    - ★ ⇒ Download a lot of data
    - ★ ⇔ `java.lang.OutOfMemoryError`
    - ★ ⇒ User must slice the sky, ...
    - ★ but no simple VO protocol allowing full-sky download by non-overlapping regions



# X-Match within the VO

## Using VO tools

- No dedicated VO protocol
- Topcat
  - ▶ Use the VO Registry and download a VOTable
  - ▶ Perform one Cone Search by row
    - ★ good for Vizier's stats...
    - ★ ...but saturates the server when multi-threaded!
- Topcat / Aladin
  - ▶ Use the VO Registry and download VOTables
  - ▶ Perform the x-match locally
    - ★ Tycho2 vs 2MASS
    - ★ ⇒ Download a lot of data
    - ★ ⇔ `java.lang.OutOfMemoryError`
    - ★ ⇒ User must slice the sky, ...
    - ★ but no simple VO protocol allowing full-sky download by non-overlapping regions



# X-Match within the VO

## Using VO tools

- No dedicated VO protocol
- Topcat
  - ▶ Use the VO Registry and download a VOTable
  - ▶ Perform one Cone Search by row
    - ★ good for VizieR's stats...
    - ★ ...but saturates the server when multi-threaded!
- Topcat / Aladin
  - ▶ Use the VO Registry and download VOTables
  - ▶ Perform the x-match locally
    - ★ Tycho2 vs 2MASS
    - ★ ⇒ Download a lot of data
    - ★ ⇔ `java.lang.OutOfMemoryError`
    - ★ ⇒ User must slice the sky, ...
    - ★ but no simple VO protocol allowing full-sky download by non-overlapping regions



# X-Match within the VO

TAP for small/medium x-matches

- Full VO x-match: TAP is the only way
- TAP (Table Access Protocol)
  - ▶ ADQL: Astronomical Data Query Language
  - ▶ UWS: Universal Worker Service

- Example

```
SELECT TOP 1000 *
FROM twomass AS a
JOIN tycho2 AS b
ON 1=CONTAINS(POINT('ICRS', a.ra, a.dec),
              CIRCLE('ICRS', b.ra, b.dec, 5./3600.))
```

- ESA GAIA access: TAP, TAP, TAP!

# X-Match within the VO

TAP for small/medium x-matches

- Full VO x-match: TAP is the only way
- TAP (Table Access Protocol)
  - ▶ ADQL: Astronomical Data Query Language
  - ▶ UWS: Universal Worker Service
- Example

```
SELECT TOP 1000 *  
FROM twomass AS a  
JOIN tycho2 AS b  
ON 1=CONTAINS(POINT('ICRS', a.ra, a.dec),  
              CIRCLE('ICRS', b.ra, b.dec, 5./3600.))
```

- ESA GAIA access: TAP, TAP, TAP!

# X-Match using TAP Vizier

TAP for small/medium x-matches

- CDS TAP Service **benchmark** (Gilles Landais)
  - ▶ **No output (simple count(\*))**
  - ▶ Before optimization (Q3C / H3C)
    - ★ Hipparcos Main (118 k) vs 2MASS (470 M): 14 min
    - ★ Tycho2 (2 M) vs 2MASS (470 M): 48 min
    - ★ 2 large catalogues: > 1 day
  - ▶ After optimization (H3C)
    - ★ Hipparcos Main (118 k) vs 2MASS (470 M): 6.5 min
    - ★ Tycho2 (2 M) vs 2MASS (470 M): 11.5 min
    - ★ GSC ACT (25 M) vs 2MASS (470 M): 1 h 30 min
- Pro: SQL flexibility
- Limit: all known TAP implementations on top of a SGBD
  - ▶ PostgreSQL (CDS, GAVO, ESAC): no multi-threading
  - ▶ Possible performances issues
  - ▶ SGBD fitted for large catalogues x-match?



# X-Match using TAP Vizier

TAP for small/medium x-matches

- CDS TAP Service **benchmark** (Gilles Landais)
  - ▶ **No output (simple count(\*))**
  - ▶ Before optimization (Q3C / H3C)
    - ★ Hipparcos Main (118 k) vs 2MASS (470 M): 14 min
    - ★ Tycho2 (2 M) vs 2MASS (470 M): 48 min
    - ★ 2 large catalogues: > 1 day
  - ▶ After optimization (H3C)
    - ★ Hipparcos Main (118 k) vs 2MASS (470 M): 6.5 min
    - ★ Tycho2 (2 M) vs 2MASS (470 M): 11.5 min
    - ★ GSC ACT (25 M) vs 2MASS (470 M): 1 h 30 min
- Pro: SQL flexibility
- Limit: all known TAP implementations on top of a SGBD
  - ▶ PostgreSQL (CDS, GAVO, ESAC): no multi-threading
  - ▶ Possible performances issues
  - ▶ SGBD fitted for large catalogues x-match?

# The CDS X-Match service

## Web interface

- Released in 2011
- General purpose service (any table having positions)  $\Rightarrow$  no probabilities
- **UWS** (Grégory Mantelet's library)
- **Very efficient** basic xmatch designed for the biggest available catalogues
  - ▶ 2 catalogues at the same time
  - ▶ **Simple** (few options)
  - ▶ E.g. SDSS DR9 vs 2MASS at 2" done in 15 min
    - ★ xmatching: 5 min (50 M links)
    - ★ building result file: 10 min (14 GB)
    - ★ running on a **single** server
- >20 G links computed in 2015, e.g.
  - ▶ NOMAD (1.1 G) vs USNOA2 (0.5 G)
  - ▶ 750 M links, 117 GB, 1h28

CDS X-Match Service

X-match Tables management Documentation

### Choose tables to cross-match

VizieR SIMBAD My store X 2MASS VizieR SIMBAD My store

**SIMBAD astronomical database**  
7,144,748 objects with position

**2MASS All-Sky Catalog of Point Sources (Cutri+2003)**  
470,992,970 rows

Hide options

**Cross-match criteria**

By position  
Radius:  arcsec

By position including error  
Sigma:  (completeness: 99.73 %)  
Max. distance:  arcsec

**Cross-match area**

All sky

Cone  
Center:   
Radius:  deg

Healpix cell (ICRS, NESTED scheme)  
Nside:   
Index:

**Begin the X-Match**

### Visualize and manage your cross-match jobs

List of X-match jobs

| Table 1        | Table 2 | Options | Begin | S |
|----------------|---------|---------|-------|---|
| No job in list |         |         |       |   |

# The CDS X-Match service

## Web interface

- Released in 2011
- General purpose service (any table having positions)  $\Rightarrow$  no probabilities
- **UWS** (Grégory Mantelet's library)
- **Very efficient** basic xmatch designed for the biggest available catalogues
  - ▶ 2 catalogues at the same time
  - ▶ **Simple** (few options)
  - ▶ E.g. SDSS DR9 vs 2MASS at 2" done in 15 min
    - ★ xmatching: 5 min (50 M links)
    - ★ building result file: 10 min (14 GB)
    - ★ running on a **single** server
- >20 G links computed in 2015, e.g.
  - ▶ NOMAD (1.1 G) vs USNOA2 (0.5 G)
  - ▶ 750 M links, 117 GB, 1h28

CDS X-Match Service

Choose tables to cross-match

SIMBAD X 2MASS

SIMBAD astronomical database  
7,144,748 objects with position

2MASS All-Sky Catalog of Point Sources (Cutri+2003)  
470,992,970 rows

Hide options

Cross-match criteria

By position

Radius:  arcsec

By position including error

Sigma:  (completeness: 99.73 %)

Max. distance:  arcsec

Cross-match area

All sky

Cone

Center:

Radius:  deg

Healpix cell (ICRS, NESTED scheme)

Nside:

Index:

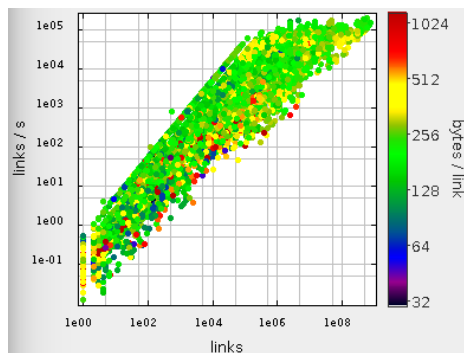
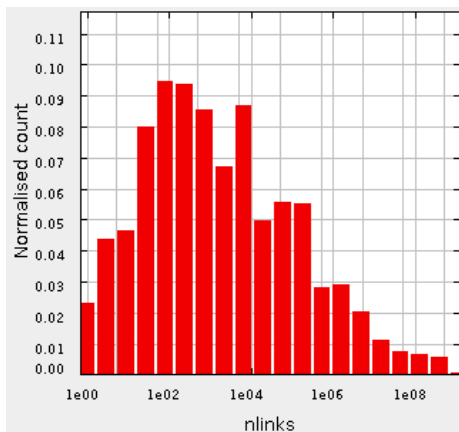
Begin the X-Match

Visualize and manage your cross-match jobs

List of X-match jobs

| Table 1        | Table 2 | Options | Begin | S |
|----------------|---------|---------|-------|---|
| No job in list |         |         |       |   |

# The CDS X-Match service



# The CDS X-Match service

HTTP API

- Released in 2013
- Uses **DALI**
- Synchronous programmatic access
- Available from
  - ▶ Web Browser, wget, curl, ...
  - ▶ TOPCAT, Portal MAST, Astropy
- Output limited to 2 M rows
- 2015: 3 G sources submitted (8 M / day)

The screenshot shows the 'CDS Upload X-Match' web interface. At the top, there is a navigation bar with 'Interop' and 'Help' menus and a toolbar with various icons. Below this is a window title bar for 'CDS Upload X-Match' with 'Window', 'Search', and 'Help' menus. The main content area is divided into three sections:

- Remote Table:** A dropdown menu is set to 'IPHAS2'. Below it, the following details are displayed:
  - Name: II/321/iphas2
  - Alias: IPHAS2
  - Description: IPHAS DR2 (218,991,524 sources)
  - Row Count: 218,991,524
  - Coverage: 0.052042644 (order 6)
- Local Table:** An 'Input Table' dropdown is set to '2: testPrec.csv'. Below it, two columns are defined:
  - RA column: 'raDeg' with units 'degrees' and a limit of '(2000)'
  - Dec column: 'deDeg' with units 'degrees' and a limit of '(2000)'
- Match Parameters:** A 'Radius' field is set to '1.0' with units 'arcsec'. A 'Find mode' dropdown is set to 'All'. A 'Rename columns' dropdown is set to 'Duplicates' and a 'Suffix' field contains 'x'. A 'Block size' dropdown is set to '50000'.

At the bottom of the interface, there are 'Go' and 'Stop' buttons.

# The CDS X-Match service

## HTTP API

- Released in 2013
- Uses **DALI**
- Synchronous programmatic access
- Available from
  - ▶ Web Browser, wget, curl, ...
  - ▶ TOPCAT, Portal MAST, Astropy
- Output limited to 2 M rows
- 2015: 3 G sources submitted (8 M / day)

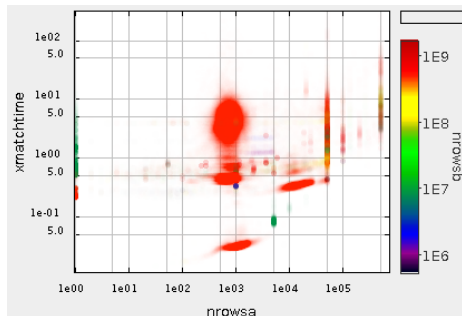
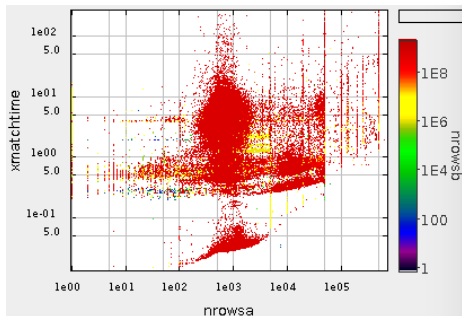
The screenshot shows a web browser window titled "CDS Upload X-Match". The interface includes a menu bar with "Window", "Search", and "Help". Below the menu are icons for home, help, and close. The main content area is divided into three sections:

- Remote Table:** A dropdown menu shows "VizieR Table ID/Alias: IPHAS2". Below it, the "Name" is "II/321/iphas2" and the "Alias" is "IPHAS2". The "Description" is "IPHAS DR2 (218,991,524 sources)", the "Row Count" is "218,991,524", and the "Coverage" is "0.052042644 (order 6)".
- Local Table:** The "Input Table" is "2: testPrec.csv". The "RA column" is "raDeg" with a unit of "degrees" and a limit of "(2000)". The "Dec column" is "deDeg" with a unit of "degrees" and a limit of "(2000)".
- Match Parameters:** The "Radius" is "1.0" with a unit of "arcsec". The "Find mode" is "All". The "Rename columns" is "Duplicates" and the "Suffix" is "\_x". The "Block size" is "50000".

At the bottom of the interface are "Go" and "Stop" buttons.

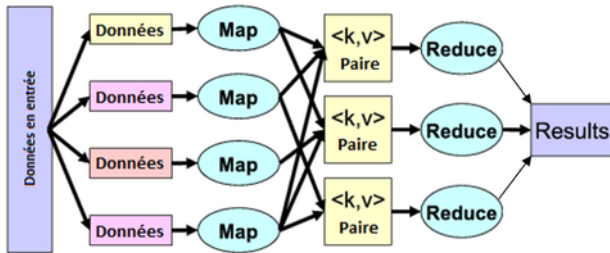
# The CDS X-Match service

HTTP API



# Test of Big Data techno

- N. Wali, A. Schaaff, F.-X. Pineau
- HADOOP / SPARK: MapReduce
- Test on SDSS DR9 / 2MASS
- Bottleneck: shuffle
  - ▶ Co-partitioning OK
  - ▶ But no way to co-locate data!! (no "block affinity groups")

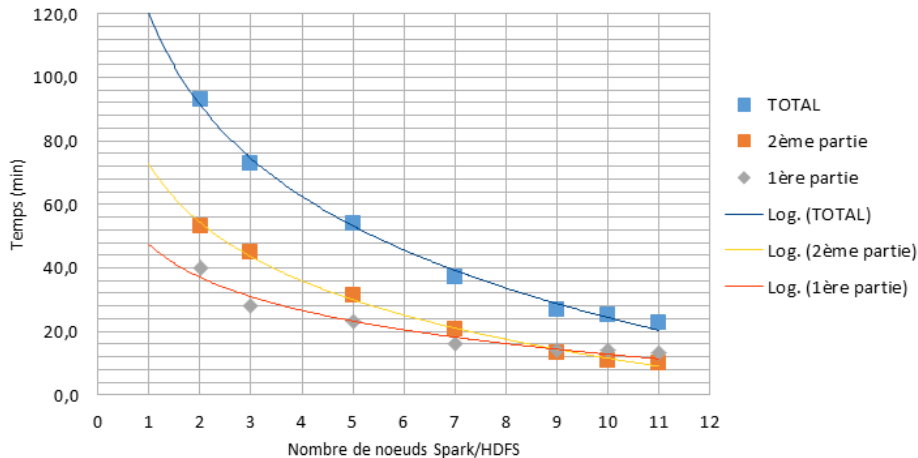


Credit: Wikipedia



# Test of Big Data techno

Temps de XMatch en fonction du nombre de noeuds



# Sommaire

- 1 Simple x-match
- 2 Multi-catalogue  $\chi$ -match & x-id

# From x-match to $\chi$ -match and x-id

- Existing tools: simple fixed radius x-match
- ARCHES (FP7): we would like identifications
  - ▶ very tricky / complex!
  - ▶ 2 (possibly) independent parts
    - ★ astrometric part (take into account positional errors)
    - ★ photometric part (not addressed here!)
    - ★  $proba_{id} \propto prior \times likelihood_{astrom} \times likelihood_{phot}$
    - ★  $proba_{id} \propto proba_{astrom} \times likelihood_{phot}$
  - ▶ SEDs  $\Rightarrow$  multi-catalogue x-match

# Candidate selection

## Method

- Steps to probabilistic positional x-match
  - ▶ Make simplifying assumptions
  - ▶ Select candidates: select and group together sources possibly being various detections of a same real source
    - ★ Need for a selection criterion
  - ▶ Make hypothesis: are the sources really from a same real sources or from different real sources?
  - ▶ For each hypothesis:
    - ★ derive the associated *likelihood*
    - ★ derive the associated *prior*
  - ▶ Compute astrometry based probabilities

# Candidate selection

## Simplifying assumptions

- Radical simplifying assumptions:
  - ▶ No proper motions
  - ▶ No blending
  - ▶ No clustering (density of sources = Poisson law)
  - ▶ No systematic offsets
  - ▶ You can trust positional uncertainties provided in catalogues

# Candidate selection

## Selection criterion

How to select a group of  $n$  sources from  $n$  distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - ▶  $H_0$  (null hypothesis): all  $n$  sources are from the same real source
  - ▶  $H_1 = \bar{H}_0$  (alternative hypothesis): at least one source (out of  $n$ ) is spurious
- User input:  $\gamma$ , the probability to accept  $H_0$  while it is true
  - ▶  $\gamma$  (I call it completeness) is called *true negative rate*
  - ▶ we usually fix  $\gamma = 0.9973$  (99.73%, value of the  $3\sigma$  rule in 1 dimensional pb)
  - ▶  $\Leftrightarrow$  fixing the **type I error** = 0.027% = proba to reject null hypothesis while it is true
  - ▶ we (theoretically) miss 27/10 000 real association
- The criterion used is simply a  $\chi^2$  test of  $2(n - 1)$  degrees of freedom

# Candidate selection

## Selection criterion

How to select a group of  $n$  sources from  $n$  distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - ▶  $H_0$  (null hypothesis): all  $n$  sources are from the same real source
  - ▶  $H_1 = \bar{H}_0$  (alternative hypothesis): at least one source (out of  $n$ ) is spurious
- User input:  $\gamma$ , the probability to accept  $H_0$  while it is true
  - ▶  $\gamma$  (I call it completeness) is called *true negative rate*
  - ▶ we usually fix  $\gamma = 0.9973$  (99.73%, value of the  $3\sigma$  rule in 1 dimensional pb)
  - ▶  $\Leftrightarrow$  fixing the **type I error** = 0.027% = proba to reject null hypothesis while it is true
  - ▶ we (theoretically) miss 27/10 000 real association
- The criterion used is simply a  $\chi^2$  test of  $2(n-1)$  degrees of freedom

# Candidate selection

## Selection criterion

How to select a group of  $n$  sources from  $n$  distinct catalogues as possibly being various observations of a same actual source?

- *Statistical hypothesis testing*
  - ▶  $H_0$  (null hypothesis): all  $n$  sources are from the same real source
  - ▶  $H_1 = \bar{H}_0$  (alternative hypothesis): at least one source (out of  $n$ ) is spurious
- User input:  $\gamma$ , the probability to accept  $H_0$  while it is true
  - ▶  $\gamma$  (I call it completeness) is called *true negative rate*
  - ▶ we usually fix  $\gamma = 0.9973$  (99.73%, value of the  $3\sigma$  rule in 1 dimensional pb)
  - ▶  $\Leftrightarrow$  fixing the **type I error** = 0.027% = proba to reject null hypothesis while it is true
  - ▶ we (theoretically) miss 27/10 000 real association
- **The criterion used is simply a  $\chi^2$  test of  $2(n - 1)$  degrees of freedom**

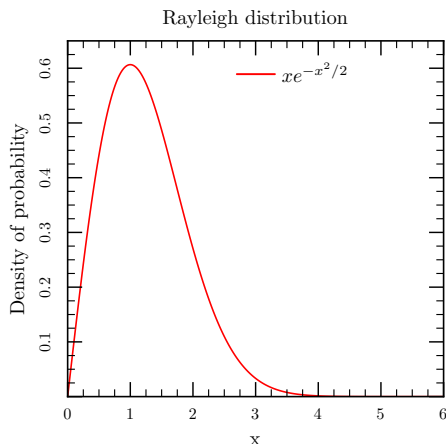


# Candidate selection

## Classical 2 catalogues case

- For real associations, i.e. when  $H_0$  is true
  - ▶ The distribution of normalized distances is a Rayleigh distribution of scale  $\sigma = 1$

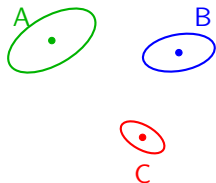
$$r \stackrel{H_0}{\sim} \text{Rayleigh}$$



# Candidate selection

## Iterative form

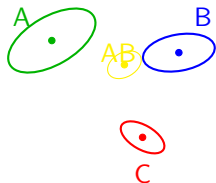
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successives and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

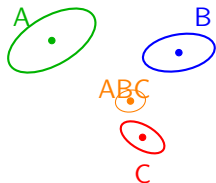
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successives and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

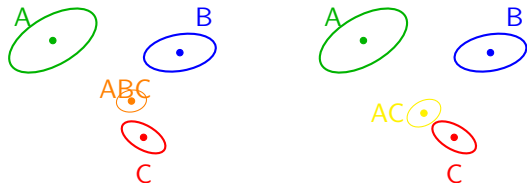
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successive and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

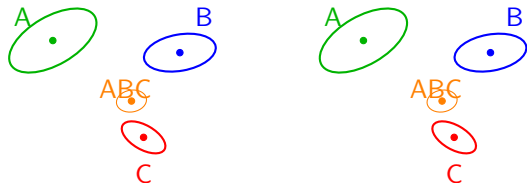
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successives and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

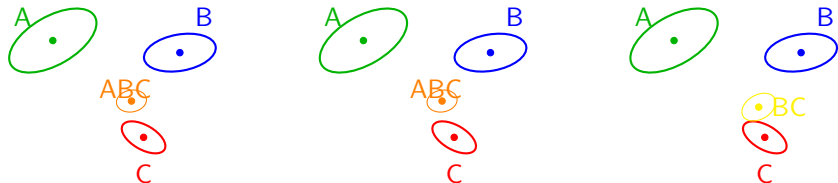
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successive and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

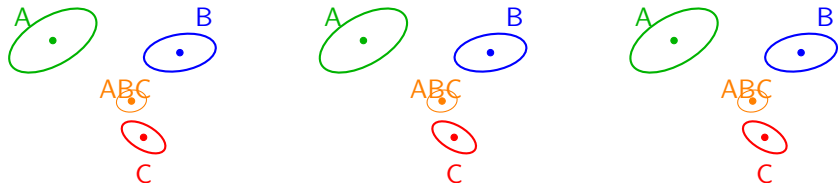
- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successive and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)



# Candidate selection

## Iterative form

- $\chi_{dof=2(n-1)} \Leftrightarrow (n-1)\chi_{dof=2}$
- $\Rightarrow$  we can perform  $(n-1)$  successive and iteratives xmatches
- At each step, we use for next position and next error ellipse
  - ▶ the weighted mean position of the previous xmatch (MLE)
  - ▶ the error on the weighted mean position of the previous xmatch
- Result independant of the xmatch order (for INNER JOIN!)





# Making hypotheses

For 2 and 3 catalogues

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

A  
•  
B

- For 3 catalogues
  - ▶ 5 hypothesis

A  
•  
B • C

# Making hypotheses

For 2 and 3 catalogues

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

A  
•  
B

- For 3 catalogues
  - ▶ 5 hypothesis

A  
•  
B • C

# Making hypotheses

For 2 and 3 catalogues

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



# Making hypotheses

For 2 and 3 catalogues

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
- ★ A\_B\_C



# Making hypotheses

For 2 and 3 catalogues

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

A  
•  
B  
•

- For 3 catalogues

- ▶ 5 hypothesis

- ★ ABC ( $H_0$ )
- ★ AB\_C
- ★ AC\_B
- ★ A\_BC
- ★ A\_B\_C

A  
•  
B • C

# Making hypotheses

For 2 and 3 catalogues

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

- Law of total probabilities:

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

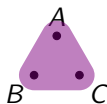
- For 2 catalogues
  - ▶ 2 hypothesis
    - ★ AB ( $H_0$ )
    - ★ A\_B

A  
•  
•  
B

- For 3 catalogues

- ▶ 5 hypothesis

- ★ ABC ( $H_0$ )
- ★ AB\_C
- ★ AC\_B
- ★ A\_BC
- ★ A\_B\_C



# Making hypotheses

For 2 and 3 catalogues

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

- Law of total probabilities:

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

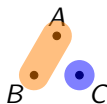
- For 2 catalogues
  - ▶ 2 hypothesis
    - ★ AB ( $H_0$ )
    - ★ A\_B

A  
•  
B  
•

- For 3 catalogues

- ▶ 5 hypothesis

- ★ ABC ( $H_0$ )
- ★ AB\_C
- ★ AC\_B
- ★ A\_BC
- ★ A\_B\_C



# Making hypotheses

For 2 and 3 catalogues

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

- Law of total probabilities:

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

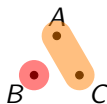
- For 2 catalogues
  - ▶ 2 hypothesis
    - ★ AB ( $H_0$ )
    - ★ A\_B

A  
•  
B  
•

- For 3 catalogues

- ▶ 5 hypothesis

- ★ ABC ( $H_0$ )
- ★ AB\_C
- ★ AC\_B
- ★ A\_BC
- ★ A\_B\_C





# Making hypotheses

For 2 and 3 catalogues

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

- Law of total probabilities:

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

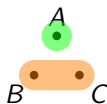
- For 2 catalogues
  - ▶ 2 hypothesis
    - ★ AB ( $H_0$ )
    - ★ A\_B

A  
•  
•  
B

- For 3 catalogues

- ▶ 5 hypothesis

- ★ ABC ( $H_0$ )
- ★ AB\_C
- ★ AC\_B
- ★ A\_BC
- ★ A\_B\_C



# Making hypotheses

For 2 and 3 catalogues

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

- Law of total probabilities:

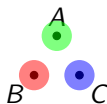
$$\sum_{i=1}^n p(H_i) = 1$$

- For 2 catalogues
  - ▶ 2 hypothesis
    - ★ AB ( $H_0$ )
    - ★ A\_B

A  
•  
B  
•

- For 3 catalogues

- ▶ 5 hypothesis
  - ★ ABC ( $H_0$ )
  - ★ AB\_C
  - ★ AC\_B
  - ★ A\_BC
  - ★ A\_B\_C



# Making hypotheses

For  $n$ -catalogues

- 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!

# Making hypotheses

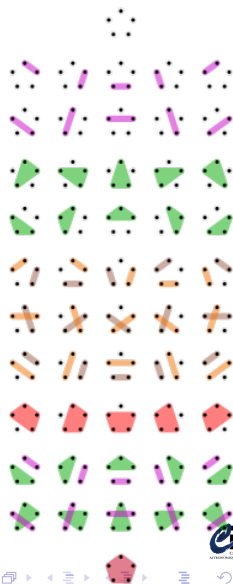
For  $n$ -catalogues

- 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!



# Making hypotheses

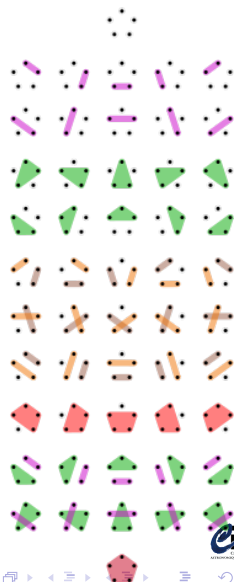
For  $n$ -catalogues

- 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 |

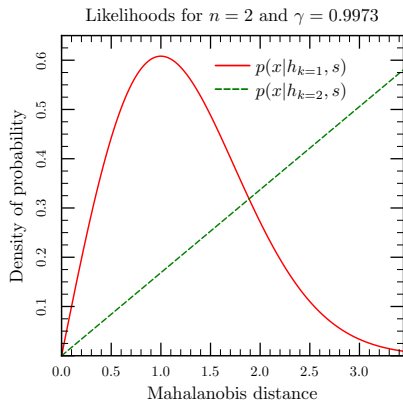
- ▶  $n$  number of catalogues
- ▶  $n=5$  catalogues  $\rightsquigarrow$  52 probabilities to be computed
- $\Rightarrow$  Combinatorial explosion!



# Likelihoods

For 2 and 3 catalogues

- Likelihood depends only on the number of “actual” source

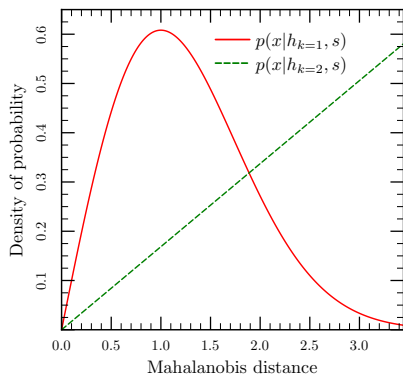


# Likelihoods

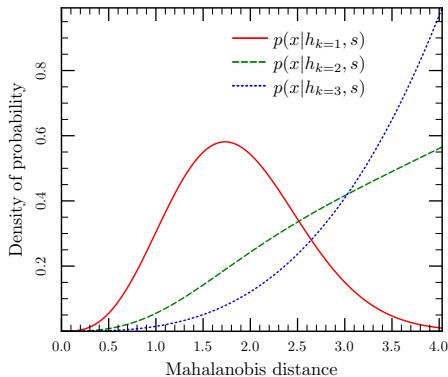
For 2 and 3 catalogues

- Likelihood depends only on the number of “actual” source

Likelihoods for  $n = 2$  and  $\gamma = 0.9973$



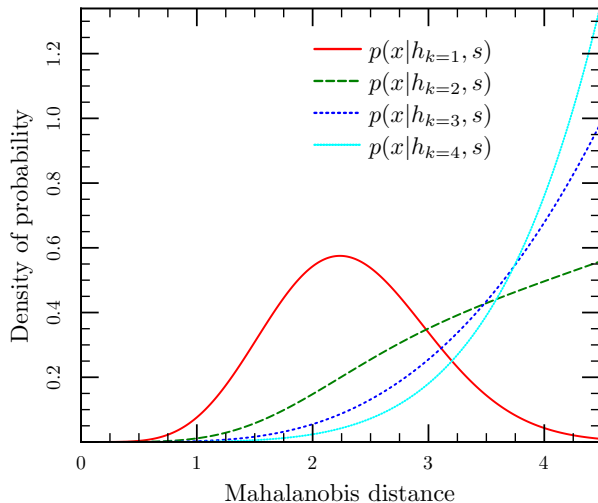
Likelihoods for  $n = 3$  and  $\gamma = 0.9973$



# Likelihoods

For 4 catalogues

Likelihoods for  $n = 4$  and  $\gamma = 0.9973$



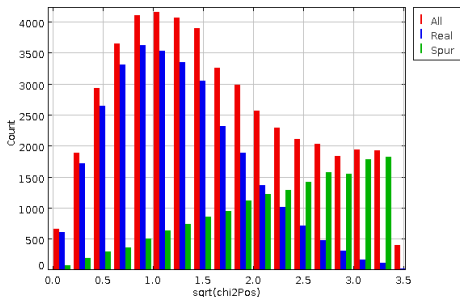


# Bayesian probabilities

For 2 catalogues

- Blue curve:  $n_{H_0} \times p(x|H_0)$
- Green curve:  $n_{H_1} \times p(x|H_1)$
- Red curve = blue + green
- For an association of given  $x$ :

$$p(H_0|x) = \frac{\text{Blue curve}(x)}{\text{Red curve}(x)}$$



- Bayes formula:

$$p(H_0|x) = \frac{p(H_0)p(x|H_0)}{p(H_0)p(x|H_0) + p(H_1)p(x|H_1)}$$

- Here priors  $p(H_0) = n_{H_0}/n_{tot}$  and  $p(H_1) = n_{H_1}/n_{tot}$

# Bayesian probabilities

For 2 catalogues

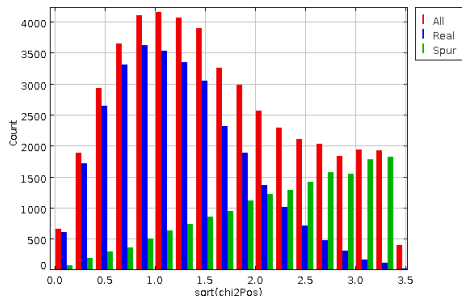
- Blue curve:  $n_{H_0} \times p(x|H_0)$
- Green curve:  $n_{H_1} \times p(x|H_1)$
- Red curve = blue + green
- For an association of given  $x$ :

$$p(H_0|x) = \frac{\text{Blue curve}(x)}{\text{Red curve}(x)}$$

- Bayes formula:

$$p(H_0|x) = \frac{p(H_0)p(x|H_0)}{p(H_0)p(x|H_0) + p(H_1)p(x|H_1)}$$

- Here priors  $p(H_0) = n_{H_0}/n_{tot}$  and  $p(H_1) = n_{H_1}/n_{tot}$



# Priors from geometrical considerations

- Common surface area of  $n$  catalogues:  $\Omega$
- Region of acceptance of the  $\chi^2$  test:
  - ▶ 2 catalogues case: ellipse
  - ▶  $n$  catalogues case:  $2(n - 1)$ -dimensional ellipsoid
- $\hat{n}_{spur} \propto$  mean volume of  $2(n - 1)$ -dimensional ellipsoid /  $\Omega^{n-1}$
- $\Rightarrow$  a way to define priors

# Test on synthetic catalogues

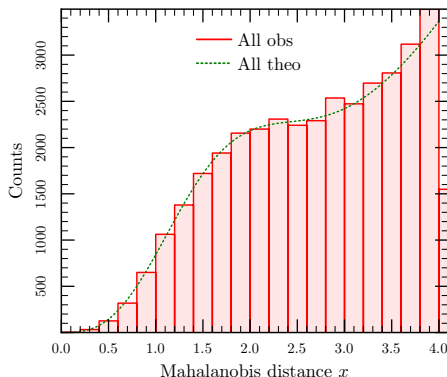
$n_A = 40\,000$   $n_B = 20\,000$   $n_C = 35\,000$

$n_{AB} = 6\,000$   $n_{AC} = 12\,000$   $n_{BC} = 18\,000$

$n_{ABC} = 10\,000$  Cone radius =  $0.42^\circ$

Err A: cte  $0.4''$ ; Err B:  $\mathcal{N}(0.75'', 0.1'')$ ; Err C: linear from  $0.8$  to  $1.2''$

Theoretical and Observed all associations



Green curve computed from geometrical considerations only!

# Test on synthetic catalogues

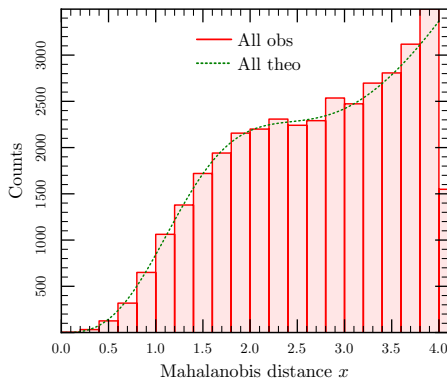
$n_A = 40\,000$   $n_B = 20\,000$   $n_C = 35\,000$

$n_{AB} = 6\,000$   $n_{AC} = 12\,000$   $n_{BC} = 18\,000$

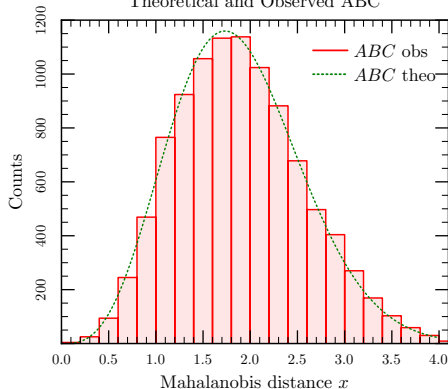
$n_{ABC} = 10\,000$  Cone radius =  $0.42^\circ$

Err A: cte  $0.4''$ ; Err B:  $\mathcal{N}(0.75'', 0.1'')$ ; Err C: linear from  $0.8$  to  $1.2''$

Theoretical and Observed all associations

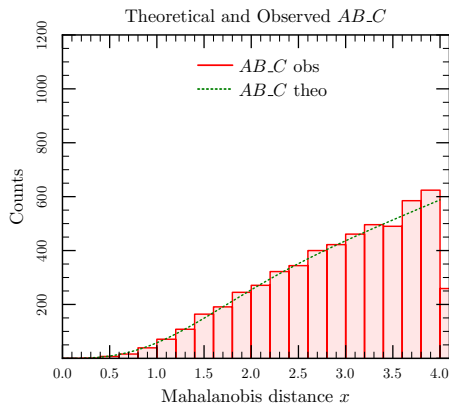
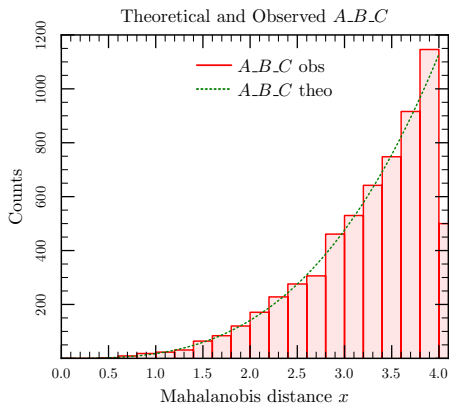


Theoretical and Observed ABC

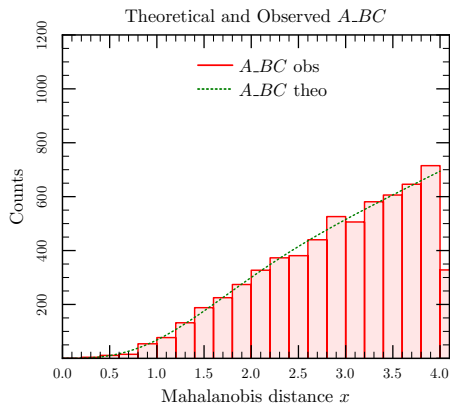
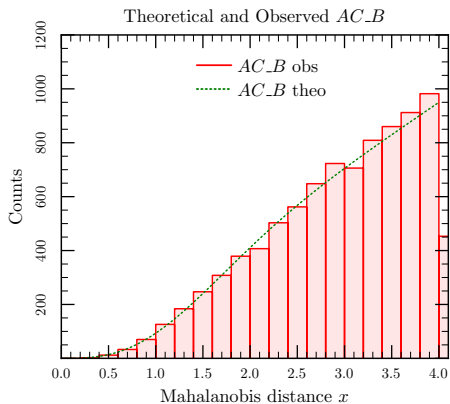


Green curve computed from geometrical considerations only!

# Test on synthetic catalogues



# Test on synthetic catalogues



## Features: various xmatch algorithms

| Algorithm         | param | #tbl | prop.mot.       | index struct. |
|-------------------|-------|------|-----------------|---------------|
| chi2 ( $\chi^2$ ) | proba | 2    | $l^1, r^2, b^3$ | M/TM-tree     |

XMatches are chainable: 1  $\chi^2$  xmatch of 4 tables = 3  $\chi^2$  xmatches of 2 tables!

4 to 11 joins ( $LIRFLIRLIRRF$ ) are supported according to the algorithm.

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.

• <sup>1</sup> mec: Minimum Enclosing Cone



## Features: various xmatch algorithms

| Algorithm         | param | #tbl | prop.mot.       | index struct. |
|-------------------|-------|------|-----------------|---------------|
| chi2 ( $\chi^2$ ) | proba | 2    | $l^1, r^2, b^3$ | M/TM-tree     |
| proba2_vx         | proba | 2    | no (?)          | M-tree        |
| proba3_vx         | proba | 3    | no (?)          | M-tree        |
| proba4_vx         | proba | 4    | no (?)          | M-tree        |
| probaN_vx         | proba | n    | no (?)          | M-tree        |

XMatches are chainable: 1  $\chi^2$  xmatch of 4 tables = 3  $\chi^2$  xmatches of 2 tables!

4 to 11 joins ( $LIRFLIRLIR'RF'$ ) are supported according to the algorithm

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.

• <sup>1</sup> mec: Minimum Enclosing Cone

## Features: various xmatch algorithms

| Algorithm         | param  | #tbl | prop.mot.       | index struct. |
|-------------------|--------|------|-----------------|---------------|
| chi2 ( $\chi^2$ ) | proba  | 2    | $l^1, r^2, b^3$ | M/TM-tree     |
| proba2_vx         | proba  | 2    | no (?)          | M-tree        |
| proba3_vx         | proba  | 3    | no (?)          | M-tree        |
| proba4_vx         | proba  | 4    | no (?)          | M-tree        |
| probaN_vx         | 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  |

XMatches are chainable: 1  $\chi^2$  xmatch of 4 tables = 3  $\chi^2$  xmatches of 2 tables!

4 to 11 joins ( $LIR\bar{F}\bar{L}\bar{R}\bar{L}'\bar{I}'\bar{R}'\bar{F}'$ ) are supported according to the algorithm.

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.

• <sup>1</sup> mec: Minimum Enclosing Cone

## Features: various xmatch algorithms

| Algorithm         | param  | #tbl | prop.mot.       | index struct. |
|-------------------|--------|------|-----------------|---------------|
| chi2 ( $\chi^2$ ) | proba  | 2    | $l^1, r^2, b^3$ | M/TM-tree     |
| proba2_vx         | proba  | 2    | no (?)          | M-tree        |
| proba3_vx         | proba  | 3    | no (?)          | M-tree        |
| proba4_vx         | proba  | 4    | no (?)          | M-tree        |
| probaN_vx         | 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 <sup>1</sup>  | dist   | n    | no (?)          | kd/M-tree     |

XMatches are chainable: 1  $\chi^2$  xmatch of 4 tables = 3  $\chi^2$  xmatches of 2 tables!

4 to 11 joins ( $LIR\bar{F}\bar{L}\bar{R}L'I'R'F'$ ) are supported according to the algorithm.

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.
- <sup>1</sup> mec: Minimum Enclosing Cone

## Features: various xmatch algorithms

| Algorithm          | param  | #tbl | prop.mot.       | index struct. |
|--------------------|--------|------|-----------------|---------------|
| chi2 ( $\chi^2$ )  | proba  | 2    | $l^1, r^2, b^3$ | M/TM-tree     |
| proba2_vx          | proba  | 2    | no (?)          | M-tree        |
| proba3_vx          | proba  | 3    | no (?)          | M-tree        |
| proba4_vx          | proba  | 4    | no (?)          | M-tree        |
| probaN_vx          | 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 <sup>1</sup>   | dist   | n    | no (?)          | kd/M-tree     |
| ext.l <sup>1</sup> | r      | 2    | no              | M-tree        |
| ext.r <sup>2</sup> | r      | 2    | no              | M-tree        |
| ext.b <sup>3</sup> | r      | 2    | no              | M-tree        |

XMatches are chainable: 1  $\chi^2$  xmatch of 4 tables = 3  $\chi^2$  xmatches of 2 tables!

4 to 11 joins ( $LIR\bar{F}\bar{L}'\bar{R}'l'R'F'$ ) are supported according to the algorithm.

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.
- <sup>1</sup> mec: Minimum Enclosing Cone

## Features: various xmatch algorithms

| Algorithm          | param  | #tbl | prop.mot.       | index struct. |
|--------------------|--------|------|-----------------|---------------|
| chi2 ( $\chi^2$ )  | proba  | 2    | $l^1, r^2, b^3$ | M/TM-tree     |
| proba2_vx          | proba  | 2    | no (?)          | M-tree        |
| proba3_vx          | proba  | 3    | no (?)          | M-tree        |
| proba4_vx          | proba  | 4    | no (?)          | M-tree        |
| probaN_vx          | 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 <sup>1</sup>   | dist   | n    | no (?)          | kd/M-tree     |
| ext.l <sup>1</sup> | r      | 2    | no              | M-tree        |
| ext.r <sup>2</sup> | r      | 2    | no              | M-tree        |
| ext.b <sup>3</sup> | r      | 2    | no              | M-tree        |
| ...                | ...    | ...  | ...             | ...           |

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

- <sup>1</sup> l: left table contains extended objects or proper motions;
- <sup>2</sup> r: right table contains extended objects or proper motions;
- <sup>3</sup> b: both left and right tables contain extended objects or proper motion.
- <sup>1</sup> mec: Minimum Enclosing Cone

# ARCHES X-match tool Web interface

## ARCHES X-MATCH TOOL Anonymous Web form



[Info about this page.](#)

### Remote directory

Upload a file:

Parcourir... Aucun fichier

### File list:

```
3xmme_uniquesources
2mass.174.10491_7.223
sdss9.174.10491_7.223
galex5ais.174.10491_7.223
```

Download Remove

### X-match script

Script examples

Chi-square xmatch of sdss/2mass in a cone, simple versio

Type, modify or copy/paste here the xmatch script to be executed:

```
1 #####
2 # Name: chi2xmatch.xml
3 # Description: Load SDSS and 2MASS data from VizieR and perform a chi-square
4 # xmatch (full join) of the two loaded tables.
5 # Remarks:
6 # - look at how we add a 0.1 arcsec systematic on SDSS positional errors
7 # - look also at how we add a new computed column to 2MASS data
8 # Input files: none
9 # Output files:
10 # - sdss_2mass.vot: result of the xmatch
11 #####
12
13 # Load SDSS DR9
14 get VizieRLoader dbname=V/139/sdss9 mode=cone center="174.10491 +7.22343" radius=12.3arcmin allcolumns
15 set pos ra=RAJ2000 dec=DEJ2000
16 set poserr type=RCD_DEC_ELLIPSE param1=e_RAJ2000 param2=e_DEJ2000
17 set cols *
18
19 # Load 2MASS data
20 get VizieRLoader dbname=II/246/out mode=cone center="174.10491 +7.22343" radius=12.3arcmin allcolumns
21 set pos ra=RAJ2000 dec=DEJ2000
22 set poserr type=ELLIPSE param1=errMaj param2=errMin param3=errPA
23 set cols *
24
25 # Perform the xmatch
26 xmatch chi2 nSten=1 nMax=1 comlatenec=0.0073 join=full
```

Submit

Result log

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

# Final words

- We plan to integrate the ARCHES engine in the CDS XMatch Service
- Multi-catalogue  $\chi$ -matches based on INNER JOINS can be performed iteratively
- No more true for e.g. LEFT JOINS (result depends on x-match order)
- Multi-catalogue x-match feasible in ADQL?
- A (evolving) standard describing how to compute probabilities (possibly several ways)??