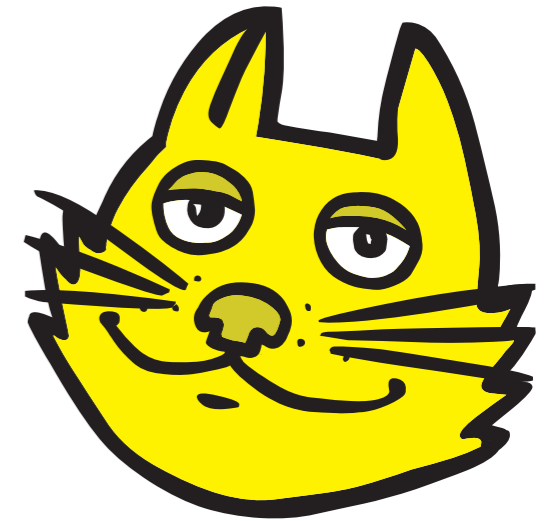# Tutorial #3: Exploring Gaia data with TOPCAT & STILTS

Mark Taylor (University of Bristol)

ESCAPE/CDS VO School
Online

23 February 2022

`$Id: tcintro.tex,v 1.3 2022/02/17 14:38:31 mbt Exp $`

TOPCAT

- What is it?
- What can it do?

Gaia mission overview

Questions?

Hands-on investigations using Gaia EDR3

- Cluster identification in velocity space
  - ▷ 2-d proper motions (M4 globular cluster)
  - ▷ 3-d velocities (NGC346 open cluster)
  - ▷ TOPCAT then STILTS

- Cross-match Gaia and HST observations
  - ▷ CDS X-Match service
  - ▷ Internal TOPCAT matching

- Gaia Herzsprung-Russell Diagram

Overview

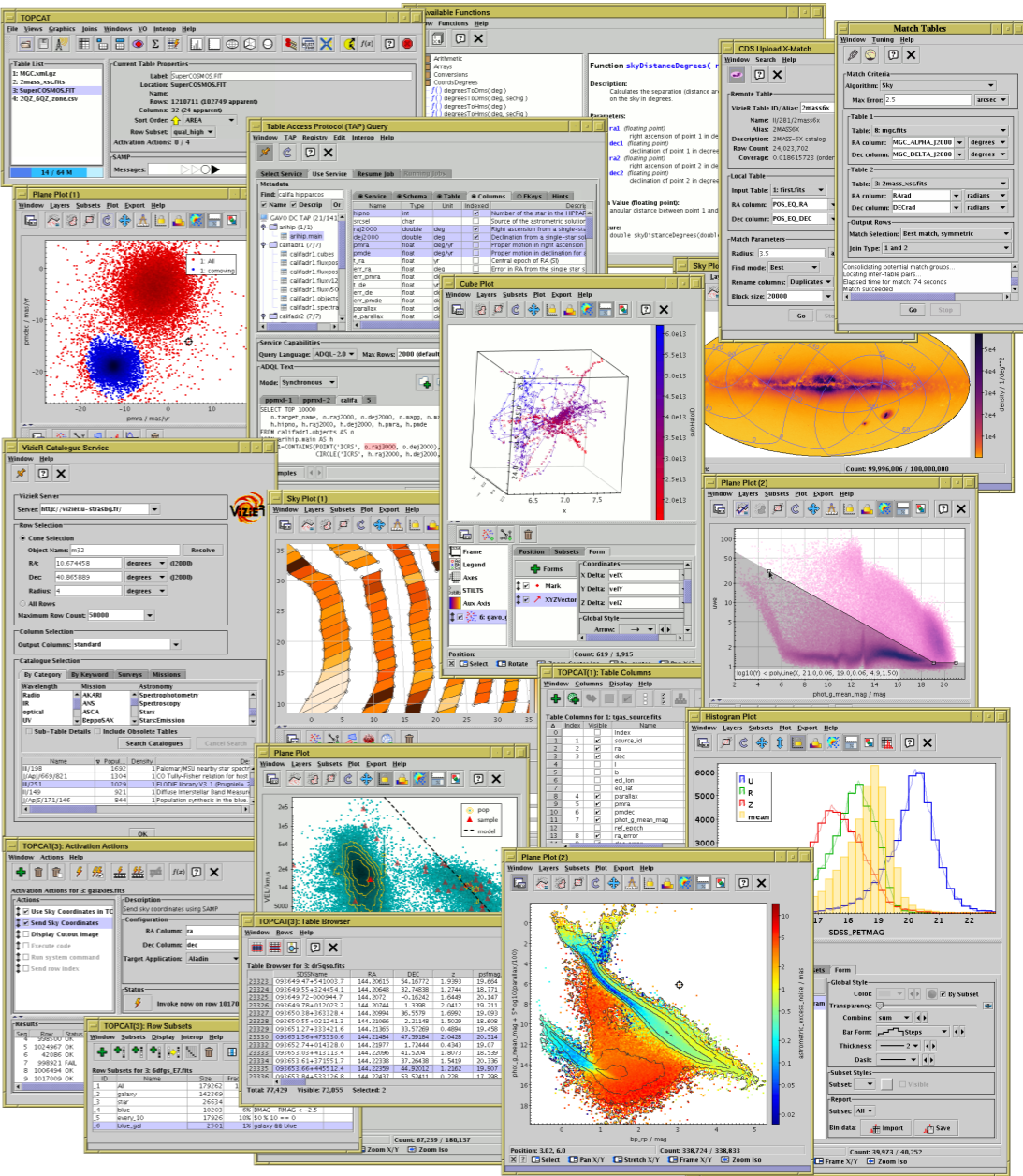*TOPCAT = Tool for OPerations on Catalogues And Tables*

"Does what you want with tables"

Suitable for:

- Interactive exploration
- Quick look at unfamilar data
- In-depth analysis

Overall aim:

- Makes table manipulation easy,
  so users can concentrate on doing science

Characteristics

Aims:

- User-friendly
  - ▷ Easy to install and run (pure Java — one download file, no library issues)
  - ▷ Easy to get started
  - ▷ Simple things fairly obvious
  - ▷ Complicated things at least well-documented
    … this does get harder as more functionality is added

- High Performance
  - ▷ Most things are fast
  - ▷ Handles quite large tables: millions of rows, hundreds of columns easily (can be much more)
  - ▷ … even on modest hardware
  - ▷ Recent/upcoming versions: better use of multi-core machines

- Do the things that astronomers need
  - ▷ Development is led by community input (mailing list, personal emails, tutorials, feature requests, bug reports…)
  - ▷ Feedback always welcome!

Capabilities

It can do:

- Read/write tables in various formats (FITS, VOTable, CSV, ...)
- View data
- View metadata
- Calculations and simple statistics (expression language)
- Visualisation (many options, interactive)
- Make/combine/display row selections in various ways (linked views)
- Crossmatching (many options)
- Access external data services (VO and others)
- Talk to other astro tools (SAMP)
- Trigger some event when a row is selected

It can't do:

- Images, spectra (it's only for tables)
- Scripting (but see STILTS)
- *Very* large tables (but see STILTS)
- All possible file formats
- Do astronomy for you

Table Data and Metadata

Table data view

Table Metadata view

Column Metadata view

Row Selections

Different ways to make single or multiple row selections:

- Select points graphically from a plot (freehand or polygon)
- Select rows from the table view
- Use an algebraic expression
- Combine existing subsets
- Receive from an external application (SAMP)

Linked views mean a selection made one way is visible in other ways

- Perform crossmatch only on items in red giant branch
- Where on the sky is this colour cut?
- Spot outliers
- Identify objects on ds9 image display

Row Highlighting

Row selection is coordinated between linked views:

- Click on row in table browser or plot
- Same row is highlighted in other plots & table browser
- Can configure external tools to highlight same object/position (SAMP)
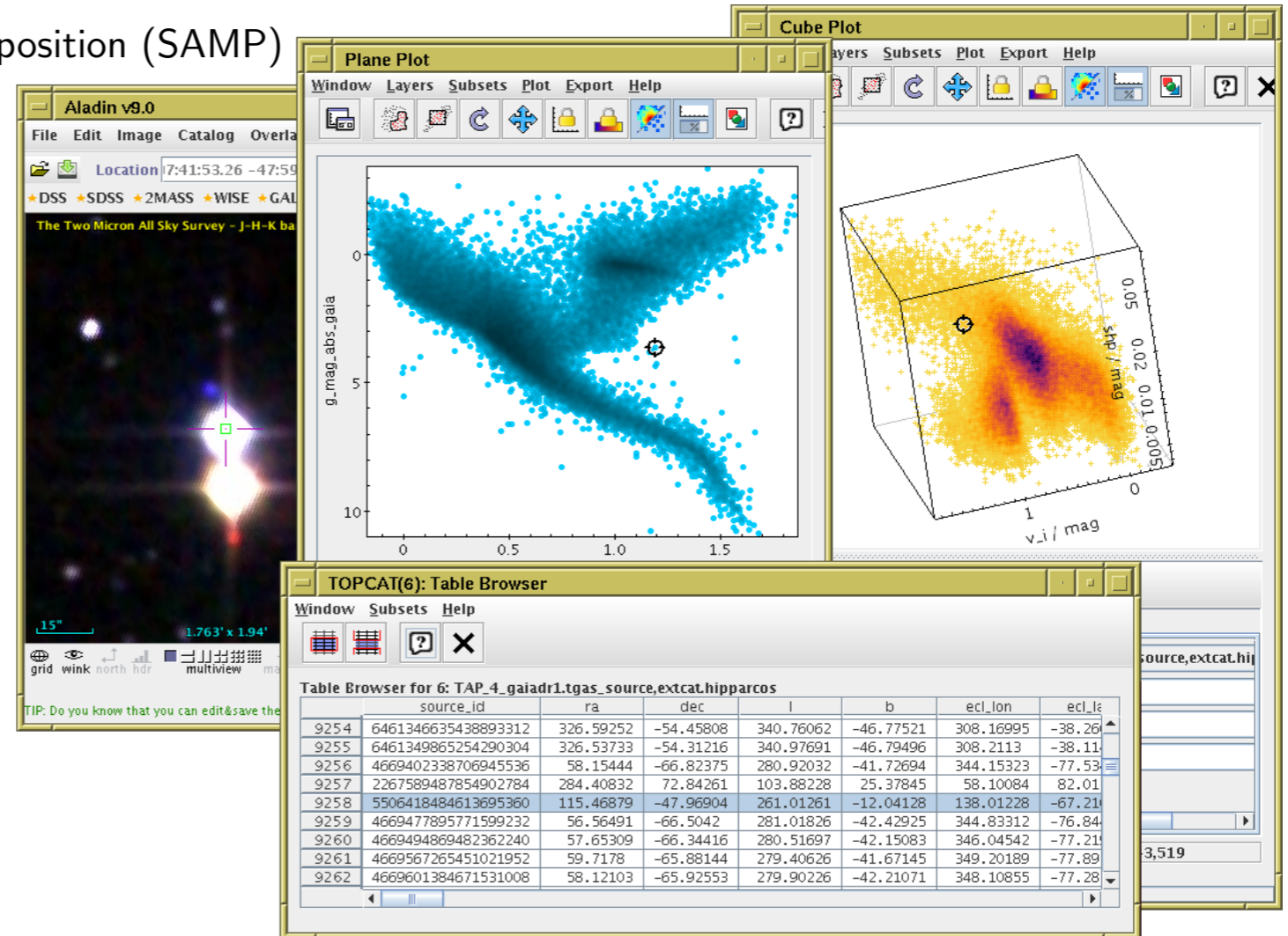
Calculations

Expression language used for creating columns, defining selections, making plots etc:

- Straightforward arithmetic syntax (C-like)
- Use column names as variables
- Standard arithmetic operators (`+`, `-`, `/`, `*`)
- Standard mathematical functions (`abs`, `max`, `round`, `sin`, `cos`, `pow`, ...)
- Conditional expressions (`q?a:b`)
- Sky coordinates (degrees, sexagesimal, sky distances)
- Astrometry (epoch propagation with/without errors, ...)
- Cosmological distances (redshift, luminosity dist, lookback time, ...)
- Fluxes (Johnson AB Magnitudes, Jansky)
- Time conversions (ISO8601, MJD, Julian, Besselian)
- ... and more (and it's extensible)

Examples:

- `mag_u - mag_g`   — *e.g. as quantity to plot*
- `janskyToAb(flux)`   — *e.g. to define new column*
- `skyDistanceDegrees(ra, dec, 14.1, -72.9) < 1.2`   — *e.g. to define row selection*

Visualisation

Very good for interactive exploration of large (or small) datasets:

- Many plot types!
  - ▷ 2d/3d scatter plots, histograms, HEALPix, density maps, error bars/ellipses, vectors, lines, quantiles, text labels, contours, KDEs, analytic functions, spectrograms, ...

- Many options!
  - ▷ Colour, colour maps, shading mode, weighting, marker shape/size line style, sky projection, sky system, coordinate grid, axis labelling, smoothing, binning, ...

- Highly responsive
  - ▷ Interactive changes to options update plot immediately

- Special attention to large data sets
  - ▷ Plot arbitrarily large datasets in fixed memory
  - ▷ Represent very dense plots in comprehensible ways
  - ▷ Many options for high-dimensional visualisation

- Publication-quality output?
  - ▷ Export to PDF, EPS, PNG, SVG, ...
  - ▷ Optional LaTeX annotation
  - ▷ Script output (STILTS) for reproducibility
  - ▷ ... but not quite as good as Matplotlib/IDL/R

Visualisation: Plot Types
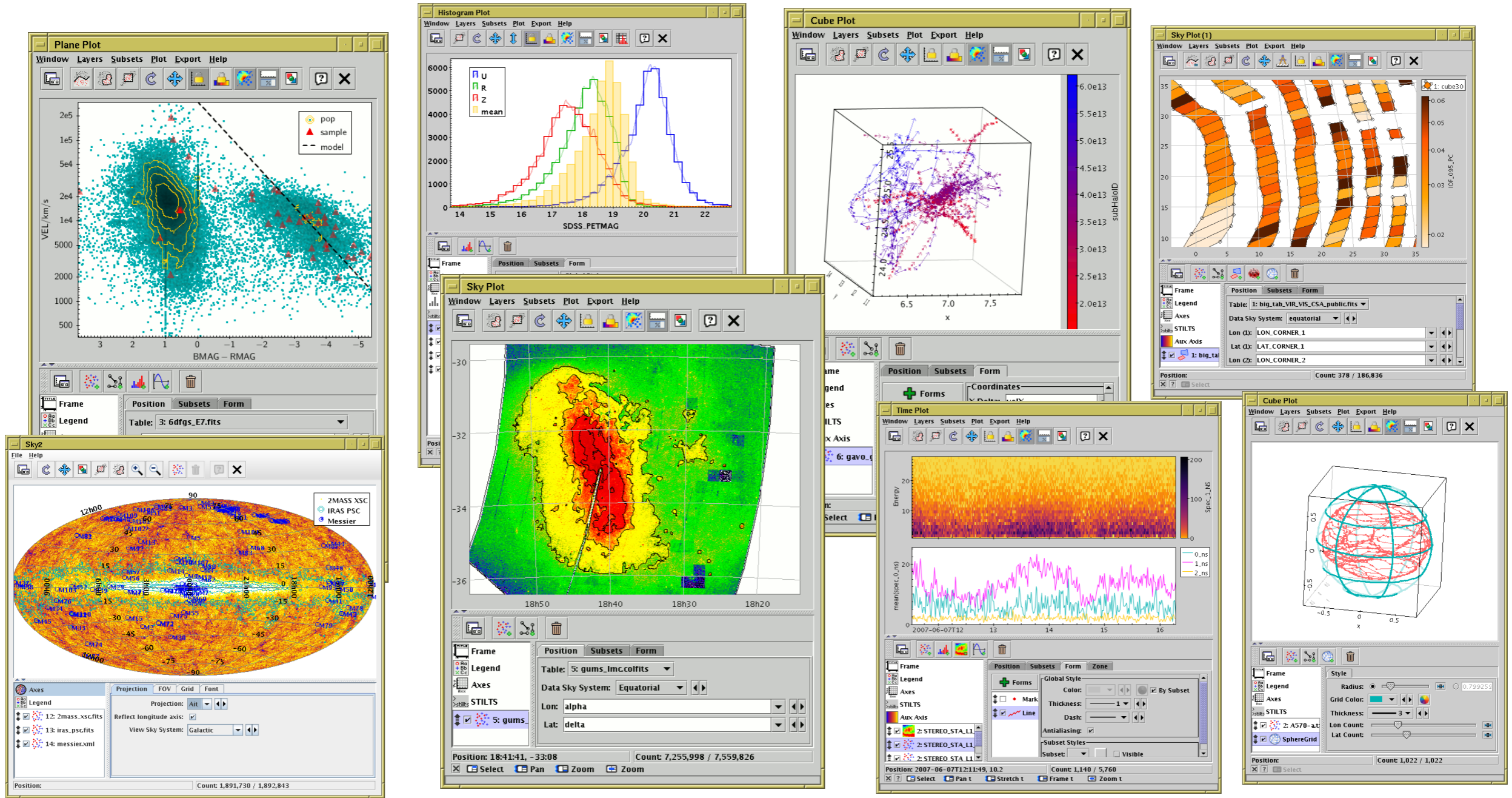
Visualisation: Dense plots
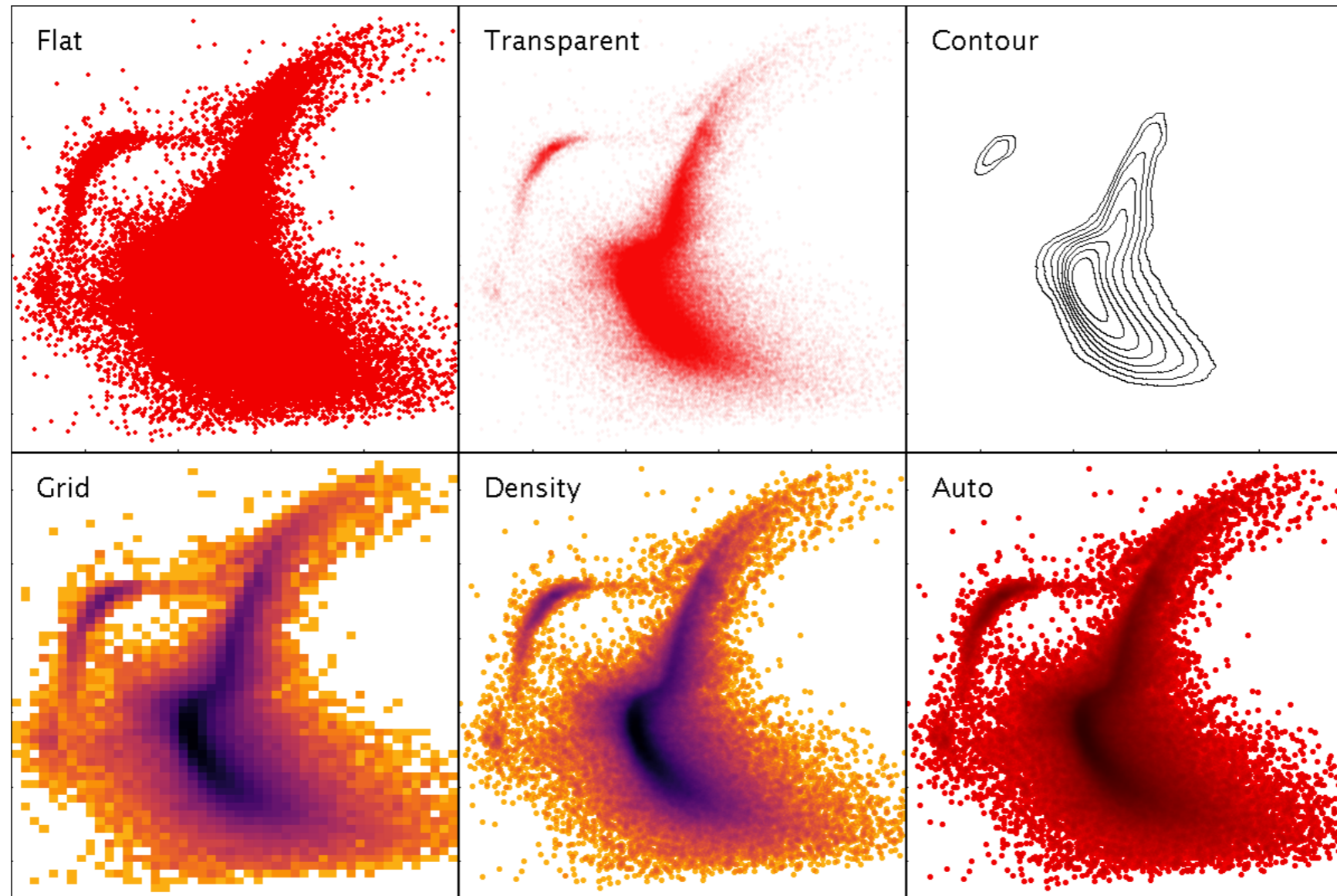
Different options for shading scatter-plot data.

Crossmatching

Internal

- Both/all files loaded into TOPCAT
- Works well up to ∼1 million rows each
- Pretty fast (≤couple of minutes)
- Very flexible (sky, Cartesian, exact, 3D, ellipses, errors, combinations…)

External

- One or both tables too big to download
- Several options, with different pros and cons:
  - ▷ CDS X-Match
    (any VizieR table, sky match, fast, easy)
  - ▷ Multiple cone search
    (many tables available, sky match, slow)
  - ▷ TAP
    (few tables available, flexible, tricky)
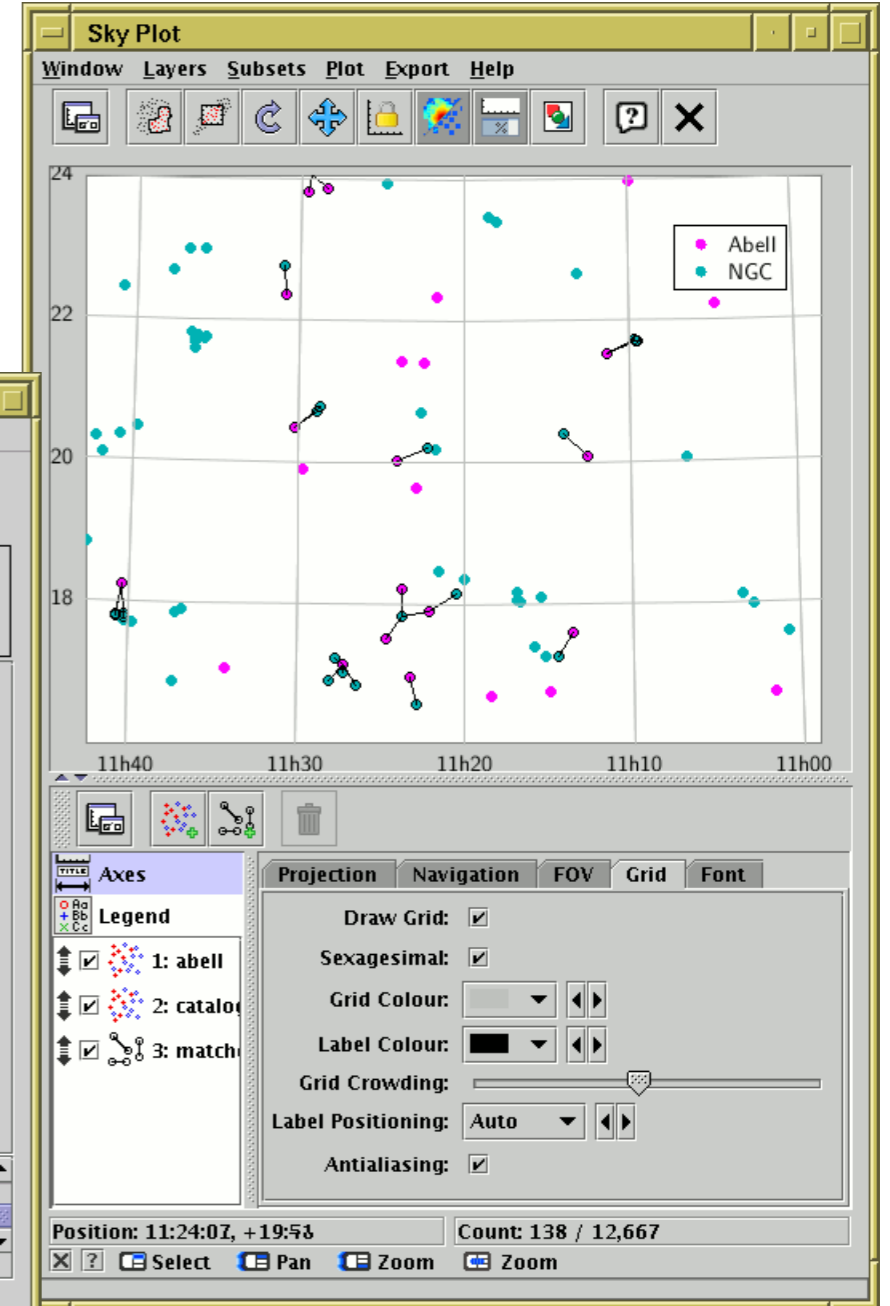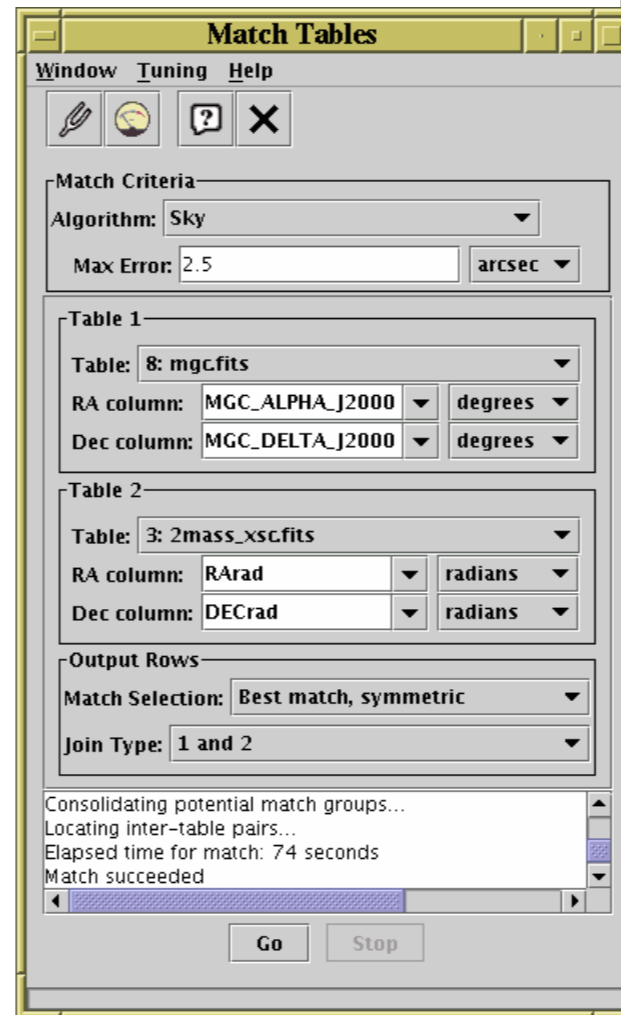
Virtual Observatory

What is the Virtual Observatory (VO)?

- *"All astro archives in your computer"*
- A set of protocols that allows software clients to talk to external data services in a uniform way
- In most cases you (the software user) don't need to understand the details, but it's under the hood making data access work

External data access from TOPCAT:

- Cone Search: positional query of remote catalogue
- Table Access Protocol (TAP): SQL-like queries against remote databases
- Simple Image Access/Simple Spectral Access: positional query of image/spectrum archives
- CDS services: Simbad, VizieR cone/all-sky, X-Match, Hips2fits
- Registry: service discovery
- SAMP: communication with other desktop/web applications

STILS

STIL Tool Set (STIL = Starlink Tables Infrastructure Library)

- Has pretty much the same capabilities as TOPCAT
- but works from the command line (also JyStilts from Jython)

TOPCAT	**STILS**
GUI	Command line
Interactive	Scriptable
Easy to use	Reproducible
Good for data exploration	Good for batch/programmed use
Exploratory phase	Production phase
few$\times 10^6$ rows	Unlimited size (for most things)

Typical usage:

- start off with TOPCAT
- maybe move on to STILTS for more specialised requirements
- TOPCAT `>stilts` STILTS control helps constructing plot commands

Further Information

- There are things I haven't mentioned!

- Full tutorial and reference documentation:
  - ▷ HTML/PDF manual on web page

    http://www.starlink.ac.uk/topcat/ (or google it)

  - ▷ **Help for Window** button [?] on every window
  - ▷ Help browser includes search tool
  - ▷ More options in Help Menu (including **Help for Window in Browser** item [?])
  - ▷ Or print out the 600-page manual

- Support:
  - ▷ **Slack** (this school)
  - ▷ email me: m.b.taylor@bristol.ac.uk
  - ▷ mailing list: topcat-user@jiscmail.ac.uk
  - ▷ All feedback and questions welcome!
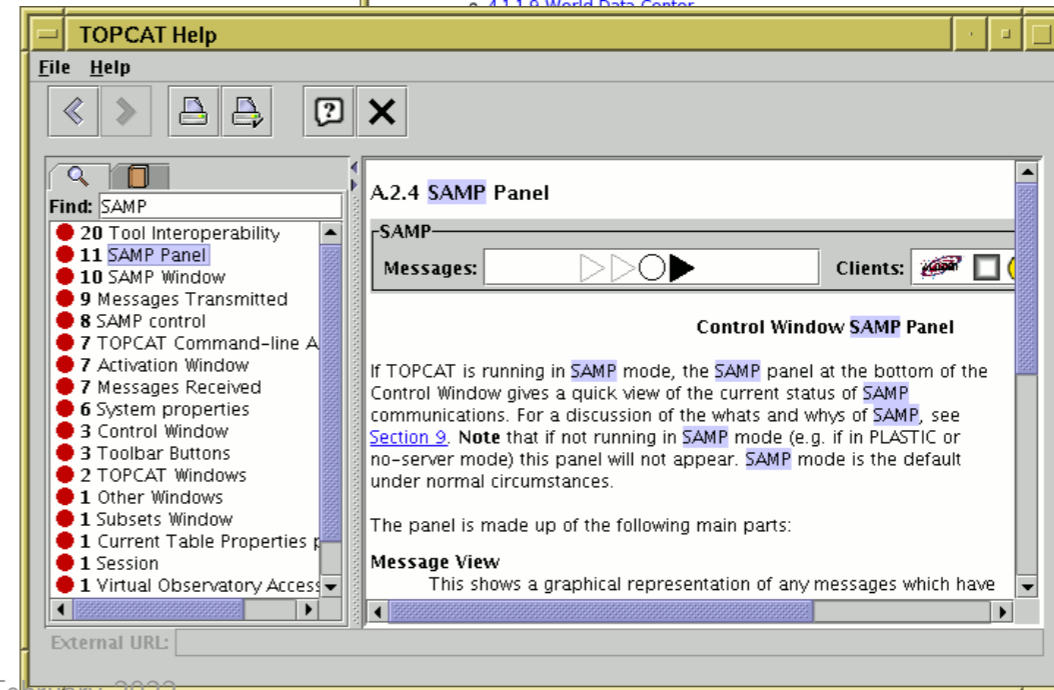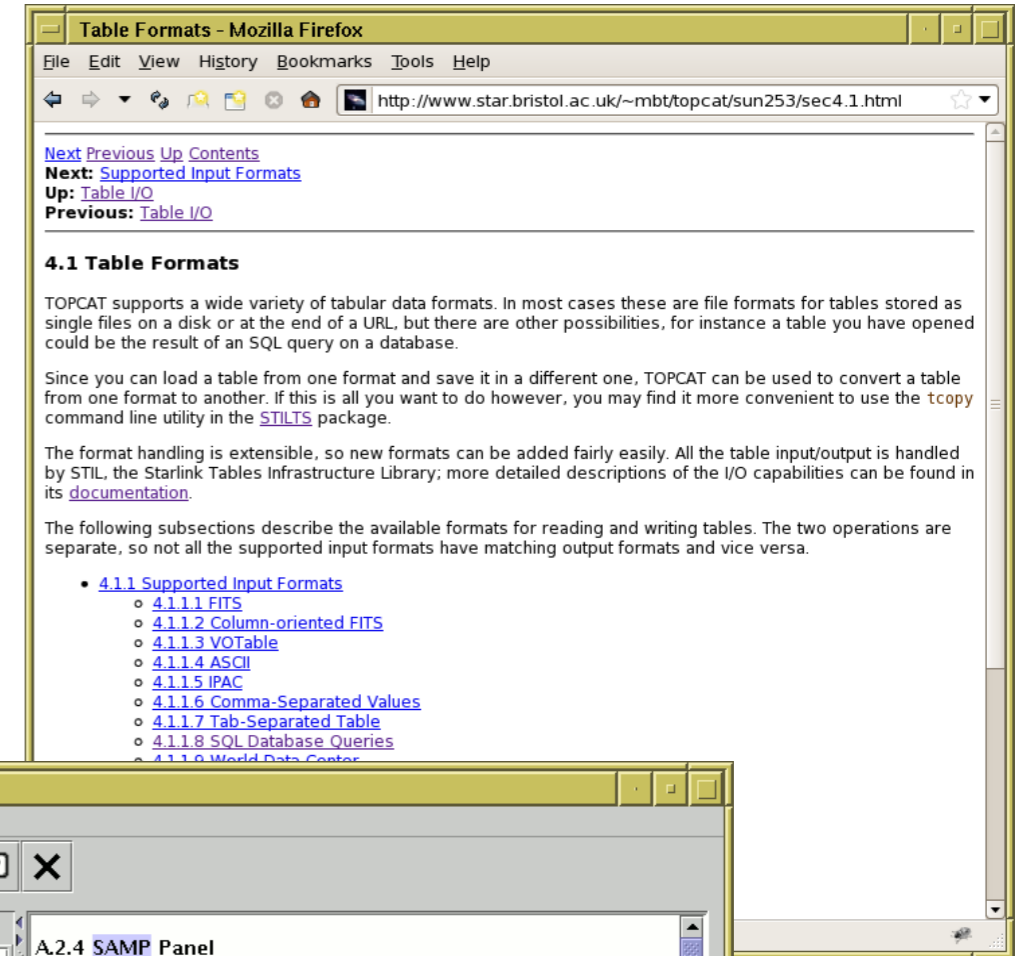
Gaia Mission

ESA astrometry satellite:

- Satellite at L2
- 5 year mission (+∼5 year extension), launched 19 Dec 2013

Aims

- Measure ∼2 billion point sources, complete to G≈20
  - ▷ mostly milky way stars, also galaxies, QSOs, SSOs, ...
  - ▷ each source observed multiple times
- Astrometry: positions, parallaxes, proper motions (10^1–$10^3\mu$as accuracy)
- Photometry: G, RP, BP bands (mmag accuracy)
- Spectrometry: radial velocities (1–15 km/s accuracy)
- Spectro-Photometry in range 330–1050 nm
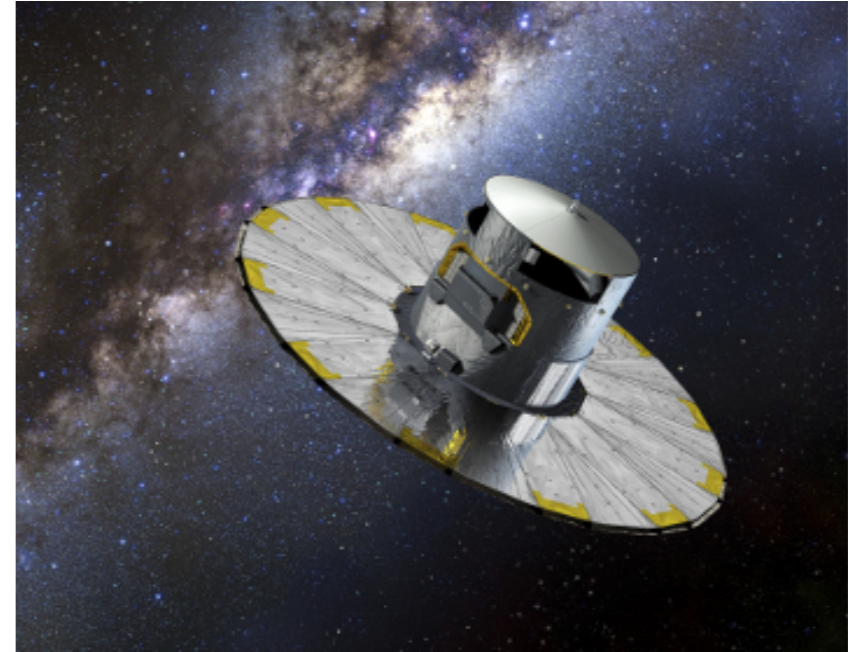
Data Releases:

- **DR1**: 14 Sept 2016 *(2-parameter astrometry, G magnitude for >1 billion sources; 5-parameter astrometry for 2 million; ...)*
- **DR2**: 25 April 2018 *(5-parameter astrometry, G, RP and BP magnitudes for >1 billion sources; R_V for 7 million; ...)*
- **EDR3**: 3 December 2020 *("Early DR3": DR2 + improved astrometry & photometry; ...)*
- **DR3**: 13 June 2022 *(EDR3 + BP/RP/RVS spectra; object classification; non-single stars; Andromeda epoch photometry; ... details)*
  - → *thousands of papers already*

Gaia EDR3 Numbers

GAIA EARLY DATA RELEASE 3

1 811 709 771
stellar positions

1 806 254 432
brightness
in white light

1 542 033 472
brightness
in blue light

1 540 770 489
colour

1 467 744 818
parallax and
proper motions

1 614 173
extragalactic
sources

1 554 997 939
brightness
in red light

#SpaceCare #ExploreFarther
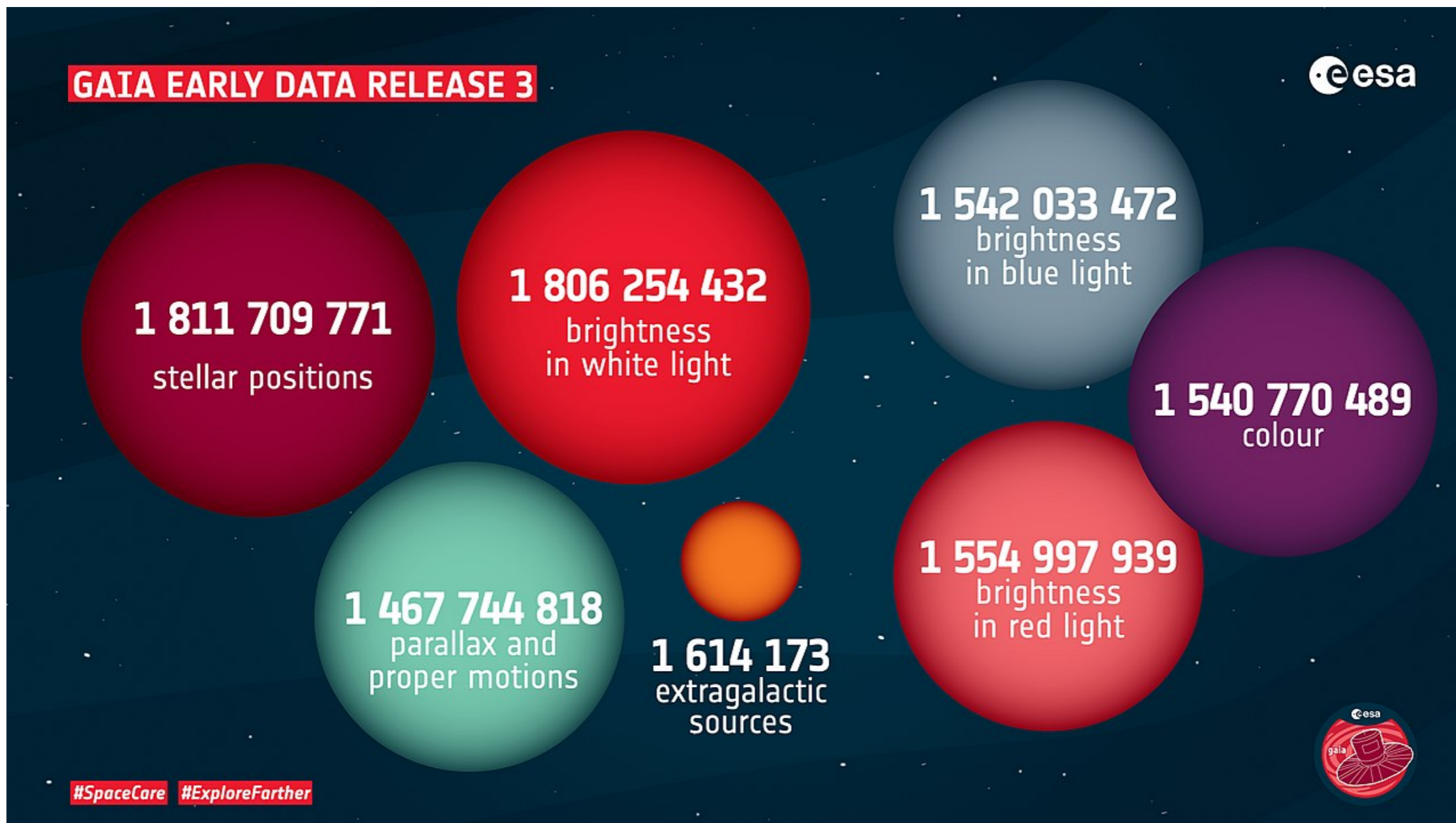
Hands-On Exercises: Topics

1. Cluster Identification #1: Messier 4 in proper motion space

   - Cone search, sky plot, subsets, histogram

2. Cluster Identification #2: Hyades in 3-D velocity space

   - TAP, expression language, 3D plot, linked views

3. Match Gaia and HST Observations for NGC346

   - VizieR download, CDS X-Match, multi-layer plots, pair match

4. Cluster Identification using STILTS

   - STILTS usage, pipelines, plotting

5. Local Herzsprung-Russell Diagram

   - Shading modes, polygon selection

Hands-On Exercises: Instructions

Exercises:

- Script is on school web site (or at **https://github.com/mbtaylor/tctuto/releases/download/gaia-edr3/tctuto.pdf**)
- For each exercise:
  - ▷ I will quickly introduce it
  - ▷ You work through the script; instructions are quite detailed, especially for the earlier exercises
  - ▷ You will have about 20 minutes for each one
  - ▷ At the end, I'll talk through it

There's quite a lot there, people work at different speeds

- If you don't finish, don't worry!
- If you finish early, explore the software and data to see what else you can do
- The main thing is to get a flavour of what can be done and how to do it

You may ask questions at any time:

- Slack? Raise hand? Zoom chat?

Exercise #1

Cluster identification #1: Messier 4 in proper motion space

- Locate Gaia EDR3 Cone Search service
- Query for sources in region of Messier 4
- Plot positions on sky
- Plot proper motions
- Create subset of comoving objects
- Create subset of background objects
- Plot proper motion vectors
- Histogram parallaxes of comoving and background objects
- Infer distance to Messier 4

Exercise #2

Cluster identification #2: Hyades in 3-D velocity space

- Locate Gaia TAP service
- Explore Gaia TAP service
- Run toy TAP query
- Run TAP query giving 6-d phase space information for nearby sources (cut'n'paste)
- Create new columns with 3-d Cartesian velocity components
- Plot sources in 3-d velocity space
- Create subset of comoving sources (Hyades)
- Examine Hyades *vs.* background sources on the sky
- Plot colour-magnitude diagram of Hyades *vs.* background sources

Note: There will be much more on ADQL/TAP in Tutorial #5 tomorrow morning

Exercise #3

Match Gaia and HST observations for NGC 346

- Download J/ApJS/166/549 catalogue from VizieR
- Crossmatch with Gaia EDR3 using CDS X-Match service
- Plot the crossmatch results
- Graphically find offset between HST and Gaia positions
- Use this to make sense of cross-match results
- Re-do crossmatch using TOPCAT internal match window

Exercise #4

Use STILTS for Messier 4 cluster identification

- Run a simple STILTS command (`calc`)
- Download M4 data from a cone search service (`cone`)
- Get used to table pipeline processing (`tpipe`)
- Obtain mean parallax for comoving objects $\rightarrow$ distance to M4
- Try some STILTS plotting (`plot2sky`)